THE STUDY OF SIGNIFICANCE OF USING MACHINE LEARNING IN DETECTION OF SUICIDAL TENDENCIES/DEPRESSION IN STUDENTS AND TEENAGERS DURING A SOCIAL MEDIA POST & VIDEO CALL

by

FARRUKH NADEEM

DISSERTATION
Presented to the Swiss School of Business and Management Geneva
In Partial Fulfillment
Of the Requirements
For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

2022

THE STUDY OF SIGNIFICANCE OF USING MACHINE LEARNING IN DETECTION OF SUICIDAL TENDENCIES/DEPRESSION IN STUDENTS AND TEENAGERS DURING A SOCIAL MEDIA POST & VIDEO CALL

by

FARRUKH NADEEM

Supervised by

MONIKA SINGH

APPROVED BY

I. Buljubasic

Dissertation chair

RECEIVED/APPROVED BY:

Renee Goldstein Osmic

Admissions Director

Dedication

I would like to dedicate this work to my parents who always wanted to see myself doing some research. I am very thankful to my mentor and SSBM that at this age I was able to even think about doing it.

ABSTRACT

THE STUDY OF SIGNIFICANCE OF USING MACHINE LEARNING IN DETECTION OF SUICIDAL TENDENCIES/DEPRESSION IN STUDENTS AND TEENAGERS DURING A SOCIAL MEDIA POST & VIDEO CALL

Farrukh Nadeem
2022

Summary:

The provided text discusses the application of machine learning techniques for suicide detection and prevention. Several approaches are explored, including:

- **Direct Detection**: Using machine learning to analyze text from social media posts and other online sources to identify individuals who may be at risk of suicide.
- **Indirect Detection**: Utilizing computer vision techniques to detect potential suicide attempts, such as depressed face or even detecting hanging, through video surveillance. This will be covered in the future scope of this research.

The limitations and challenges of these methods are also highlighted, including data quality, model accuracy, and ethical considerations.

Key findings and future directions:

- **Machine learning shows promise**: It can effectively detect suicide ideation and potential attempts.
- **Data quality is crucial**: High-quality, annotated data is essential for training accurate models.
- **Human intervention is necessary**: While machine learning can provide valuable insights, human judgment is still needed to make informed decisions.

Future research: Focus on improving model accuracy, addressing ethical concerns, and exploring new techniques like deep learning and natural language processing. Utilize the RestAPI end points during video calls can play a crucial role in detecting any negative thoughts. Introducing other vital metrics such as BMI, Blood pressure etc. may further enhance the accuracy of our model's findings.

Overall, the text emphasizes the potential of machine learning to revolutionize suicide prevention but acknowledges the need for continued research and development.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1: Introduction

Chapter 2: Literature Review

Chapter 3: Methodology

Chapter 4: Results and Discussion

LIST OF FIGURES

Chapter 1: Introduction

Chapter 2: Literature Review

Chapter 3: Methodology

Chapter 4: Results and Discussion

Chapter 5: Summary and Recommendations

CHAPTER I:
INTRODUCTION

1.1 Introduction

Suicide and depression have increasingly been recognised as critical public health concerns among adolescents and young adults, both at the global level and within India. The World Health Organization (2021) reports that more than 700,000 people die by suicide each year, making it one of the leading causes of mortality among individuals aged 15–29. The situation in India is particularly alarming, with official statistics recording over 13,000 student suicides in 2022—equivalent to one every 40 minutes—and reflecting a 64 per cent increase over the past decade (National Crime Records Bureau, 2023). This trend not only surpasses the overall national suicide growth rate but also highlights the vulnerability of young learners to mental health challenges. Globally, more than one billion people are affected by mental health conditions, a burden that is further intensified by the rapid expansion of digital and social media platforms. While these technologies provide opportunities for connection and self-expression, they also expose young people to risks such as cyberbullying, online harassment, and unhealthy social comparison, which can exacerbate anxiety, depression, and diminished self-esteem (Livingstone et al., 2021).

Although awareness of suicide and depression among young people has grown, stigma continues to act as a major obstacle to prevention and treatment. Many students remain reluctant to seek professional help due to fears of being judged, punished, or misunderstood, which often results in underreporting and social withdrawal (Patel et al., 2018). The persistence of cultural taboos surrounding mental health, combined with limited awareness and inadequate institutional support, further intensifies the problem (Kumar and Singh, 2020). In the Indian context, the shortage of trained mental health professionals and the unequal distribution of services—particularly in rural and economically disadvantaged regions—leave large sections of vulnerable youth without adequate care (National Mental Health Survey of India, 2016). These systemic challenges underscore the need for holistic and culturally responsive strategies that integrate both social and technological approaches. Emerging applications of machine learning, for example, hold promise for early identification of at-risk individuals and timely intervention (Sharma and Arora, 2021). By combining technological innovation with socially sensitive frameworks, this research seeks to contribute to more effective and sustainable models of suicide prevention among students and adolescents.

The English word *suicide* originates from the Latin term *suicidium*, meaning “to kill oneself.” Today, suicide is recognised as one of the leading causes of mortality worldwide, ranking among the top ten causes of death on a global scale (World Health Organization, 2019). Current estimates suggest a global suicide rate of approximately 16 per 100,000 people, with millions of deaths attributed to suicide each year. It is therefore regarded as one of the most pressing social and public health challenges of contemporary society. Suicidal behaviour is shaped by a complex interplay of personal and social factors, including exposure to trauma, physical or mental illness, social isolation, hopelessness, and anxiety (Patel et al., 2018).

According to the World Health Organization (2019), more than 700,000 individuals died by suicide in 2019, representing roughly one in every 100 deaths worldwide. In response to this alarming trend, the WHO has issued updated guidelines to assist countries in strengthening suicide prevention and care (World Health Organization, 2021). While suicide attempts are more common than completed suicides, particularly among adolescents, suicidal ideation—defined as thoughts or plans to end one's life—remains a critical predictor of risk. Such ideation may be fleeting or persistent, and in some cases involves detailed planning, rehearsal, or unsuccessful attempts (Klonsky, May and Saffer, 2016). A significant proportion of young people report experiencing suicidal thoughts, underscoring the importance of early detection and timely intervention as strategies to reduce suicide-related fatalities.

Suicide has emerged as a critical global public health concern, with multiple life circumstances—such as stress, depression, unemployment, failure, and social pessimism—contributing to suicidal behaviour (Patel et al., 2018). The Centers for Disease Control and Prevention (CDC) in the United States has highlighted not only the human cost but also the substantial economic burden associated with suicide. It is estimated that suicide and suicide attempts result in an annual loss of approximately 70 billion US dollars in combined medical expenses and lost productivity (Centers for Disease Control and Prevention, 2022). These figures underscore the urgent need for governments to prioritise suicide prevention strategies, with a focus on early detection, timely intervention, and comprehensive support systems to safeguard both individual wellbeing and broader social stability (World Health Organization, 2021).

Suicide remains a significant public health challenge, particularly among young people. In the United States, it is the second leading cause of death for individuals aged 10–34 (National Institute of Mental Health, 2023). Suicidal ideation, often referred to as suicidal thoughts, involves the serious consideration of ending one's own life. Such thoughts may be triggered by a range of emotional and situational factors, including anger, guilt, anxiety, depression, or experiences of failure and disappointment (Klonsky, May and Saffer, 2016). Although most individuals who experience suicidal ideation do not proceed to attempt suicide, the persistence of such thoughts can contribute to long-term psychological distress and increase the risk of eventual suicide if appropriate interventions are not provided (Safai, 2024). Early recognition and treatment of suicidal ideation are therefore essential components of effective suicide prevention strategies.

With the help of doctors and medicines, suicidal thoughts may be controlled. However, most individuals who struggle with suicidal thoughts opt not to get medical help due to feelings of shame. As an alternative, today many individuals now announce their suicide plans or their depressive thoughts on social media.

The most effective method of avoiding suicide may be the early identification of warning symptoms or risk factors, because mental illness may be recognized and treated. An effective method of preventing suicide among those who are very inclined to attempt it is to facilitate communication between these people and professionals such as therapists or social workers. However, there may be diagnostic delays with this approach, which might prove deadly for the patient. Identifying suicidal thoughts early is key to effective suicide prevention. The content of user postings may be analyzed for signs of suicide ideation.

A person's level of psychological, emotional, and social health all has a role in his mental health and, in turn, his thoughts, feelings, and actions. The way a person handles stress, his decision-making abilities, his interactions with other people, his relationships & productivity during the workplace, and every other aspect of his life, are all influenced by his level of happiness. Anyone of any age, gender, race, or socioeconomic status is vulnerable to developing a mental health issue.

A wide range of mental health conditions are recognised globally, including depression, mood and behavioural disorders, suicidal and self-harming behaviour, bipolar disorder, borderline personality disorder, dissociative disorders, anxiety, panic attacks, hypomania, paranoia, phobias, schizophrenia, and disturbances related to eating and sleeping patterns (Patel et al., 2018). The prevalence of these conditions has risen sharply in recent decades, influenced by multiple factors such as genetic predisposition, family history, adverse life experiences, chronic stress, poor physical health, and the pressures of fast-paced modern living. Social changes, including the decline of extended family structures and reduced emotional support networks, have further exacerbated vulnerability to mental illness (Kumar and Singh, 2020).

Mental and substance use disorders were estimated to affect approximately 15.5 per cent of the global population in 2016 (Ahrnsbrak et al., 2016), with India accounting for nearly 15 per cent of its population. In Europe, the World Health Organization (2021) reports that around 27 per cent of adults have experienced a mental health condition at some point in their lives. Similarly, in the United States, one in six adults—equivalent to 18.3 per cent—experience a mental illness each year (National Institute of Mental Health, 2023). Suicide rates reflect these challenges: in the United States, suicide ranks as the tenth leading cause of death overall and the second among young adults aged 15–24, while India records one of the highest suicide rates globally in the 15–29 age group (Radhakrishnan and Andrade, 2012). The increasing prevalence of depression among adolescents, particularly in the context of widespread social media use, has intensified concerns, with reports indicating that in India, a student dies by suicide every hour, often after expressing suicidal thoughts online.

There have been many shocking cases of individuals broadcasting their suicides online, and all of them have been discovered to have suffered from depression in the past. We believe that if there had been a real-time system to identify such incidents and inform the individuals in the person's social network, many lives may have been spared. When compared to matters of physical health, conversations regarding one's mental well-being tend to take a back seat. People with mental health problems often go undiagnosed for a long time, even by themselves. Lack of education and information about them, as well as societal shame, contribute to their neglect. This may increase the patient's propensity toward suicide or self-harming behavior by adding to his existing mental anguish, suffering, and frustration.

Machine learning is a branch of artificial intelligence that focuses on algorithms capable of improving performance through experience (Mitchell, 1997). Recent advancements in probabilistic modeling have significantly enhanced the ability to handle uncertainty in data-driven systems (Bishop, 2006). These developments have laid the foundation for modern applications in pattern recognition, predictive analytics, and decision-making (Mitchell, 1997; Bishop, 2006).

Given the severe consequences of untreated mental health conditions, early identification and intervention are of critical importance. Research indicates that untreated mental illness can, in some cases, be associated with violent behaviour, underscoring the need for timely support and care (Fazel and Grann, 2006). Early recognition of symptoms, combined with emotional support from family and friends and access to professional treatment, can significantly improve recovery outcomes and help individuals reintegrate into daily life (Patel et al., 2018). Although it is not always possible to diagnose mental illness with certainty at an early stage, observable changes in behaviour, emotions, and mood often serve as warning signs. These may include persistent negative thoughts, mood instability, destructive or self-harming tendencies, and atypical emotional outbursts such as anger, frustration, or aggression (American Psychiatric Association, 2013). Recognising these indicators at an early stage is therefore essential for effective prevention and intervention.

In clinical practice, psychiatrists and psychologists routinely observe behavioural and emotional cues during consultations, using structured questions to assess a patient’s mental state, thoughts, and feelings. Diagnostic insights are often derived from the individual’s responses and self-reported experiences (American Psychiatric Association, 2013). However, in the contemporary digital era, social media has become a dominant medium through which individuals express their emotions and share personal reflections, often more openly than in face-to-face interactions (Naslund et al., 2016). This shift suggests that continuous monitoring of online emotional expressions could provide valuable indicators of emerging mental health concerns. Long-term analysis of such digital footprints may therefore support the early detection of psychological distress and facilitate timely intervention (Chancellor and De Choudhury, 2020).

Suicide continues to be one of the leading causes of mortality worldwide, consistently ranking among the top ten causes of death across regions. The World Health Organization (2021) estimates that more than one million people die by suicide each year, equating to a global mortality rate of approximately 16 per 100,000 individuals, or one death every 40 seconds. Younger populations have increasingly emerged as the most vulnerable groups, surpassing older men in many countries. Mental and substance use disorders remain a major contributing factor, affecting around 15.5 per cent of the global population in 2016, including nearly 15 per cent of India's population (Ahrnsbrak et al., 2016). In Europe, more than a quarter of adults are estimated to have experienced a mental health condition at some point in their lives (World Health Organization, 2021), while in the United States, one in six adults (18.3 per cent) experiences a mental illness annually (National Institute of Mental Health, 2023). Suicide is the second leading cause of death among Americans aged 15–24 and the tenth overall, while India records one of the highest suicide rates among young people aged 15–29 (Radhakrishnan and Andrade, 2012).

It is estimated that over 90 per cent of suicides are linked to underlying mental health conditions, particularly depression and substance misuse (Ferrari et al., 2014). Recognising this, many governments have adopted suicide prevention strategies to address the crisis (National Institute of Mental Health, 2022). Anxiety and depression, which are increasingly prevalent in both developed and emerging economies, remain central concerns. Without adequate treatment, individuals with severe mental illness are at heightened risk of suicidal ideation and attempts. The proliferation of harmful online content, including cyberbullying, cyberstalking, and misinformation, further exacerbates psychological distress and can have devastating consequences (Chancellor and De Choudhury, 2020).

Studies indicate that victims of cyberbullying are particularly vulnerable to suicidal behaviour (Hinduja and Patchin, 2019). Suicide, however, is a multifaceted phenomenon shaped by a combination of personal, social, and historical factors. The American Foundation for Suicide Prevention (2022) categorises risk factors into three domains: individual circumstances, societal influences, and historical precedents. Ferrari et al. (2014) highlight mental health and substance use disorders as key determinants, while O'Connor and Nock (2014) emphasise psychological vulnerabilities such as personality traits, cognitive patterns, and adverse life experiences.

The digital era has also given rise to concerning social phenomena, including online communities that normalise self-harm and encourage copycat suicides. A notable example is the so-called “Blue Whale Challenge,” which emerged in 2016 and was associated with a series of self-harm tasks culminating in suicide, reportedly linked to multiple deaths worldwide (Mukhra, Baryah and Krishan, 2017). At the same time, the internet has become a space where individuals increasingly disclose suicidal thoughts, often under the veil of anonymity. This has created opportunities for the use of social media mining and machine learning techniques to detect early signs of suicidal ideation and support prevention efforts (Sharma and Arora, 2021).

While suicidal ideation reflects psychological suffering, its predictive value as a screening tool remains limited. A meta-analysis by McHugh et al. (2019) demonstrated that suicidal thoughts alone are insufficient to reliably forecast suicide attempts. Nevertheless, early recognition of ideation by social workers and mental health professionals, combined with proactive engagement, can help alleviate distress and reduce risk. Suicide is ultimately the result of multiple interacting factors, and researchers continue to employ psychological assessments, clinical investigations, and structured questionnaires to better identify and understand suicidal thoughts.

Since the rise of social media, increasing numbers of individuals have used digital platforms such as blogs, microblogs, and online forums to disclose suicidal thoughts and emotional distress (Park, Cha and Cha, 2012; Bathina et al., 2020). Adolescents, in particular, often turn to these spaces to articulate suicidal ideation, seek advice from online communities, or, in extreme cases, participate in suicide pacts. The anonymity afforded by online communication enables users to share personal anxieties and pressures more openly than in face-to-face interactions (Naslund et al., 2016). These online disclosures present both risks and opportunities: while they may normalise harmful behaviours, they also provide valuable data for the early detection and prevention of suicide.

Among social media platforms, X (formerly Twitter) has become one of the most widely studied environments for sentiment analysis and mental health research. Analyses of user posts have demonstrated the potential to identify conditions such as depression and anxiety, thereby informing innovative approaches to mental health care (Chancellor and De Choudhury, 2020). Advances in machine learning and Natural Language Processing (NLP) have further enabled the development of predictive models capable of detecting suicidal ideation in online content, offering promising avenues for early intervention and suicide prevention strategies (Sharma and Arora, 2021).

Numerous single set feature groups were extracted by various researchers, including Linguistic Inquiry Word Count (LIWC) (Coppersmith, Dredze, Harman, & Hollingshead, 2015; Maupomé & Meurs, 2017), Bag-of-Words (Nadeem, 2016; Paul, Jandhyala, & Basu, Aug. 2018), and Ngrams (Wongkoblap & Curcin, 2017; Benton & Hovy. Latent Dirichlet Allocation (an LDA), or Sept. 2018), for identifying depression in user communications. Other research (Resnik et at., 2015; Preotiuc-Pietro et al., 2015; Nguyen et al., Jul. 2014; Schwartz et al., 2014) used different machine learning approaches to compare the performance of these particular features.

Recent research has increasingly focused on the integration of multiple linguistic and computational features to enhance the accuracy of detecting mental health conditions through social media data. Tsugawa et al. (2015), for instance, demonstrated that combining N-Gram features with the Linguistic Inquiry and Word Count (LIWC) tool produced higher detection accuracy than using either feature independently. Similarly, Wolohan et al. (2018) improved performance by combining Bag-of-Words, Latent Dirichlet Allocation (LDA), Term Frequency–Inverse Document Frequency (TF-IDF), and Convolutional Neural Networks (CNNs), supported by advanced text preprocessing. Tadesse et al. (2019) also highlighted the benefits of hybrid approaches, showing that while a single bi-gram feature with a Support Vector Machine (SVM) achieved 80 per cent accuracy, combining LIWC, LDA, and bi-grams with a Multilayer Perceptron (MLP) increased accuracy to 91 per cent.

This growing body of literature demonstrates the potential of composite features in improving the early detection of mental illness through social media analysis. Studies by Benamara et al. (2018), Song, You and Park (2018), Cacheda et al. (2019), and Tadesse et al. (2019) further reinforce the value of hybrid models, showing that integrating multiple computational and linguistic features can significantly enhance predictive performance in mental health monitoring.

Using deep neural networks like CNNs, RNNs, LSTMs, etc. for different supervised and unsupervised tasks, the most recent research in the field of artificial intelligence and Machine Learning is producing unprecedented results. Unstructured data, such as photos, videos, and texts, are ideal for these deep learning systems to detect patterns in. The vast majority of online text and material is created by individuals who lack formal training or experience in content writing, hence it is extremely unstructured. Online postings or comments on different social media platforms such as Twitter, Facebook, etc. are the most prevalent kind of material made by the users on a regular basis. These updates often include brief text, visuals, or a combination of the three. This user's postings are a window into their thinking and a method of cutting-edge communication in the current online world.

Advances in machine learning and natural language processing (NLP) have enabled the analysis of social media content as a means of identifying early indicators of depression, suicidal ideation, and self-harming behaviour. By examining the linguistic and emotional patterns expressed in user-generated content, algorithms can decode and interpret underlying psychological states (Chancellor and De Choudhury, 2020). Emotional cues are often embedded in the frequency and nature of posts, including the use of emoticons, images, videos, and textual expressions that convey sadness, grief, disappointment, or other negative affective states (Bathina et al., 2020). Such digital traces provide valuable opportunities for the early detection of mental health concerns and the development of timely, technology-driven interventions (Sharma and Arora, 2021).

The COVID-19 pandemic, first identified in Wuhan, China in late 2019, reached India with its first confirmed case on 30 January 2020. Since then, the country has recorded millions of infections and over one million deaths, making it one of the hardest-hit nations globally (Ministry of Health and Family Welfare, 2021). While public health responses have largely concentrated on diagnosis and treatment of the virus, comparatively less attention has been directed towards its psychological consequences.

The pandemic has profoundly affected mental health by generating uncertainty about the future, disrupting social support systems, creating financial instability, and exposing individuals to illness and bereavement. These stressors have been strongly associated with heightened psychological distress, which in turn has been linked to suicidal ideation and self-destructive behaviours (Gunnell et al., 2020). Suicide remains a major global public health concern, with the World Health Organization (2019) estimating that one person dies by suicide every 40 seconds. In India, nearly 171,000 suicides were reported in 2022, representing a 4.2 per cent increase from 2021 and a 27 per cent rise compared to 2018 (Swain et al., 2021).

Both the WHO and the American Foundation for Suicide Prevention (AFSP) have identified a range of risk factors that lower the threshold for suicide, including depression, anxiety, insomnia, schizophrenia, substance misuse, bereavement, and a history of suicide attempts (AFSP, 2022; World Health Organization, 2021). A common misconception is that suicide provides a permanent solution to temporary problems. However, evidence suggests that early recognition of risk factors and timely referral for professional support can prevent many deaths. Unfortunately, the shortage of trained mental health professionals, combined with persistent stigma, continues to discourage individuals from seeking help (Patel et al., 2018).

As highlighted in previous research, the rapid growth of Social Networking Services (SNSs) has created spaces where individuals feel increasingly comfortable discussing mental health concerns without fear of stigma or reprisal (Naslund et al., 2017; Berry et al., 2017). These platforms provide opportunities to observe and analyse behavioural patterns through user-generated content, offering potential avenues to identify individuals at risk of suicidal ideation. This is particularly valuable in contexts where limited resources, inadequate screening mechanisms, and persistent stigma hinder access to traditional mental health services. Among SNSs, Twitter has emerged as one of the most widely studied platforms, functioning as a searchable repository of users' perspectives, emotions, and experiences, and thereby serving as a valuable resource for mental health research and intervention.

Scientists have scoured Twitter for insights on a wide range of topics, including e illness, emotional issues, and a history of previous suicide attempts (Faryal et al., 2021). Except for North Korea, Iran, and China, Twitter is available everywhere. People don't have to worry about being judged for what they write on Twitter, thus the content often reflects how they really feel. It's believed that every day, 500 million tweets are sent out, and that 23% of all internet users are active on Twitter for social networking (Rachna Swamy, 2021). Twitter saw the need for a tool to report suicide material, and implemented it. However, this is contingent on the perception of networked users, and hence, the true danger cannot be understood using this method.

Another place where individuals may voice their ideas online is Reddit. Subreddits are discussion forums devoted to a certain topic. There is a subreddit called "SuicideWatch (SW)" where users discuss ways they could commit themselves. The subreddit will be kept up to date by the Moderators. It is possible to save lives by being prepared for the specific linguistic characteristics that people with suicide thoughts use to convey their distressing emotions.

Even while some individuals are good at hiding their emotions, indicators like sadness, anxiety, and hyperactivity might point to a deeper depression that could lead to actual suicide attempts. In addition, with the rise of social networks, individuals are increasingly turning to these platforms to share and discuss ideas with others. They are more likely to follow through on their goal to share their sentiments with the public if they actively participate in online discussion and community forums. It has been reported that some social media users have used the anonymity provided by the platform to announce their suicide plans. Virtually no websites advocate such tragic outcomes or successfully sway users with noble justifications.

The early identification of suicidal tendencies is critical, as timely intervention can prevent loss of life and enable more effective medical and psychological treatment. Detecting suicidal ideation before it escalates into action is therefore a vital component of suicide prevention strategies (O'Connor and Nock, 2014). A range of methods has been proposed to achieve this, drawing on diverse data sources such as online activity, messaging behaviour, voice patterns, and non-verbal cues.

Textual analysis has been widely explored, with techniques including lexicon-based filtering, word cloud visualisation, Natural Language Processing (NLP), decision tree classifiers, recurrent neural networks (RNNs), and deep learning approaches such as convolutional neural networks (CNNs) and long short-term memory (LSTM) models (Tadesse et al., 2019). In addition, Support Vector Machines (SVMs) have been applied to voice and audio data, where Mel Frequency Cepstral Coefficients (MFCCs) are used to capture frequency patterns and paralinguistic features associated with depression (Al Hanai, Ghassemi and Glass, 2018).

Other machine learning approaches, including Learning Vector Quantization (LVQ), K-Nearest Neighbours (KNN), LSTM neural networks, and ensemble averaging techniques, have also been employed to improve detection accuracy. Beyond text and audio, researchers have investigated the role of facial expressions and non-verbal behaviour in identifying suicidal risk. Tools such as OpenFace 2.0, combined with human–computer interaction (HCI) frameworks and computer vision methods, have been used to analyse micro-expressions and behavioural cues that may indicate psychological distress (Baltrušaitis, Robinson and Morency, 2016).

Suicide is a multifaceted phenomenon influenced by a wide range of psychological, social, and biological factors. While not all suicides can be attributed solely to mental illness, psychiatric conditions such as major depressive disorder and bipolar disorder are strongly associated with elevated suicide risk, in some cases increasing vulnerability by more than twenty-fold (Hawton and van Heeringen, 2009). Emotional distress often precedes suicidal behaviour, with individuals perceiving suicide as a final option when they reach a psychological breaking point.

In parallel, advances in artificial intelligence (AI) and machine learning (ML) have opened new possibilities for suicide research and prevention. Machine learning, a branch of AI, enables systems to learn patterns from data and make predictions without explicit programming (Jordan and Mitchell, 2015). Within this context, ML models can be trained on historical suicide data to identify risk factors, predict trends, and estimate suicide rates across populations. The accuracy of such models depends heavily on the quality and volume of available data, as well as the choice of algorithms and feature engineering techniques.

For example, regression and classification methods are commonly employed in suicide prediction studies. Regression analysis is used to quantify the relationship between dependent and independent variables, making it a valuable tool not only in psychology and sociology but also in economics and business decision-making (Montgomery, Peck and Vining, 2012). Classification algorithms, on the other hand, allow researchers to categorise individuals into risk groups based on behavioural, demographic, or clinical features. More advanced approaches, such as neural networks and ensemble methods, have also been applied to improve predictive accuracy in suicide risk modelling (Berrouiguet et al., 2019).

When the dependent variable can be divided into two categories, logistic regression is typically the preferred analytical method. In cases where the distribution between the two categories is relatively balanced (close to 50/50), the results of logistic regression and linear regression may appear comparable (Hosmer, Lemeshow and Sturdivant, 2013). Regression analysis can incorporate both continuous and dichotomous independent variables. Categorical variables with more than two levels are often transformed into binary indicators through a process known as dummy coding, which allows them to be included in regression models (Field, 2018). Importantly, regression identifies associations between variables but does not, on its own, establish causal direction.

Classification, by contrast, is a supervised machine learning technique that organises data into predefined categories. It can be applied to both structured and unstructured datasets. The process involves training a predictive model to map input features to discrete output classes, with the primary objective of assigning new data points to the most appropriate category (James et al., 2021). Classification algorithms—such as decision trees, support vector machines, and neural networks—are widely used in predictive modelling, particularly in domains such as health informatics, finance, and social media analysis.

Difficulties with speech recognition, face recognition, handwriting recognition, document categorization, etc., are among the most frequent problems encountered in classification systems. It may include more than two categories of people, as well.

Reports indicate that in India, approximately one student dies by suicide every hour, with many young people disclosing suicidal thoughts on social media platforms prior to the act (Radhakrishnan and Andrade, 2012). Adolescents, who are among the most active users of social media, have also experienced a marked rise in depression rates globally (World Health Organization, 2021). Disturbingly, several cases have emerged in which individuals live-streamed their suicides online, with many of these individuals having a documented history of depression. These incidents highlight the urgent need for real-time monitoring systems capable of detecting suicidal expressions on social media and alerting members of the individual's network. Such interventions could provide opportunities for timely support and potentially prevent tragic outcomes (Chancellor and De Choudhury, 2020).

Compared with physical health conditions, mental health remains one of the least openly discussed topics. Individuals experiencing mental illness may not recognise their condition for extended periods, and stigma, coupled with limited awareness, often leads to neglect or dismissal of their struggles (Patel et al., 2018). This neglect exacerbates psychological suffering, frustration, and distress, increasing the likelihood of self-harm or suicidal behaviour. Early identification and treatment of mental illness are therefore critical, as delays in intervention can contribute not only to suicidal ideation but also, in some cases, to aggression or violence towards others (Fazel and Grann, 2006). Several high-profile cases, including mass shootings, have been linked to untreated mental health conditions.

Timely diagnosis, supported by emotional care from family and friends alongside professional medical assistance, can significantly improve recovery outcomes. Although it is not always possible to confirm a diagnosis at an early stage, observable behavioural and emotional changes often serve as warning signs. These may include mood swings, persistent feelings of hopelessness, destructive thoughts, or uncharacteristic emotional outbursts such as anger and frustration (American Psychiatric Association, 2013). In clinical practice, psychiatrists and psychologists rely on such cues, using structured questioning to assess mood, thoughts, and behaviour, and drawing diagnostic insights from patient responses.

In today’s digital era, individuals increasingly express themselves online rather than in face-to-face interactions. This shift suggests that affective analysis of social media posts, conducted over extended periods, may provide valuable insights into early indicators of mental health difficulties (Chancellor and De Choudhury, 2020). Recent advances in deep learning architectures—including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) models—have demonstrated exceptional capacity to detect patterns in unstructured data such as text, images, and video (LeCun, Bengio and Hinton, 2015). Given that most online content is user-generated and highly unstructured, posts on platforms such as Facebook and X (formerly Twitter) offer a unique window into users’ thoughts and emotions. Properly analysed, these posts can reveal expressions of sadness, grief, disappointment, and other negative affective states, thereby supporting the early detection of depression, suicidal ideation, and self-harming behaviours (Bathina et al., 2020).

Against this backdrop of rising mental health challenges and persistent barriers to care, it is essential to examine the underlying factors contributing to suicide and depression among students and adolescents. The following section reviews the current literature on youth mental health, the role of social media, and the evolution of suicide prevention strategies, providing the foundation for the research questions and methodological framework of this study.

Social Constraints and Disclosure Barriers

Despite the growing awareness of mental health issues, many students refrain from openly sharing their emotional struggles with parents, even during video calls. Cultural stigma surrounding mental health, fear of judgment, and academic pressure often compel adolescents to maintain a façade of well-being. Being away from home amplifies this tendency, as students seek independence and privacy while avoiding potential consequences such as restrictions or withdrawal from academic programs. These social constraints underscore the need for proactive detection mechanisms that do not rely solely on verbal disclosure, making multimodal approaches—leveraging text, audio, and visual cues—critical for early intervention.

This research introduces a novel integration of machine learning techniques with adaptive content delivery systems, extending their application beyond conventional predictive analytics. Unlike traditional approaches that focus solely on algorithmic performance, this work emphasizes interdisciplinary utility by bridging machine learning with therapeutic practices such as Neuro-Linguistic Programming (NLP). The proposed framework enables personalized positive content delivery through digital platforms, fostering mental well-being and patient engagement. By combining predictive modeling, language-driven behavioral reinforcement, and real-time adaptability, this study offers a unique contribution to both technical innovation and societal impact. Potential real-world applications include mobile health apps, voice-assisted therapy tools, and web-based platforms designed to support cognitive and emotional resilience.

1.1 Contributions of this Study

1. **Multimodal framework:** We design and empirically assess a text–audio–video fusion pipeline for early detection of suicidal ideation in adolescents, contrasting deep models (LSTM, CNN, LSTM–CNN) with classical baselines (SVM, NB, DT, RF, XGBoost).
2. **Culturally adaptive NLP:** We analyze code-mixed (Hindi–English) and idiomatic expressions and document error modes that arise in non-Western contexts, informing culturally grounded annotation and modeling.
3. **Human-in-the-loop workflow:** We specify operational protocols for counselor validation of alerts, consent flows, and escalation pathways suitable for schools/telehealth.
4. **Ethics & compliance blueprint:** We contribute a pragmatic GDPR/HIPAA compliance checklist, privacy-by-design measures, and fairness auditing guidance for deployment.
5. **Application to NLP-based therapy:** We outline an adaptive positive-content delivery loop for patient engagement, connecting risk detection to supportive interventions

1.2 Scope & Delimitations:

This study focuses on adolescents and young adults and evaluates ML models primarily using public social-media text and research-consented audio/video interactions. It does not conduct clinical diagnosis, does not replace clinical judgment, and does not implement live interventions. The results generalize to education and telehealth contexts with similar populations; extension to other demographics and languages requires re-validation.

BACKGROUND AND RELATED STUDIES

Suicide has become one of the most urgent global public health challenges of the 21st century. The World Health Organization (2021) reports that more than 700,000 people die by suicide each year, with suicide ranking as the second leading cause of death among individuals aged 15–29. Adolescents and young adults are disproportionately affected, with increasing rates of depression, anxiety, and self-harm behaviours documented across both high-income and low- and middle-income countries (Patel et al., 2018). Beyond the devastating human toll, the economic burden is substantial: the Centers for Disease Control and Prevention (2020) estimate that suicide and suicide attempts cost the United States over $70 billion annually in medical expenses and lost productivity.

The digital era has reshaped the ways in which young people communicate, express emotions, and seek support. Social media platforms such as X (formerly Twitter), Reddit, Instagram, and Facebook have become central outlets for self-expression among students and teenagers. While these platforms can foster connection, they also expose vulnerable users to cyberbullying, social comparison, and online harassment—factors strongly associated with depression and suicidal ideation (Chancellor and De Choudhury, 2020). In India, the crisis is particularly severe: the National Crime Records Bureau (2024) reported 13,044 student suicides in 2022, reflecting an annual increase of nearly 4 per cent, which is double the national suicide growth rate. A pooled 26-year analysis published in *Child and Adolescent Psychiatry and Mental Health* (2024) further confirmed the upward trajectory of youth suicides in India, projecting continued increases over the next decade.

Traditional approaches to suicide prevention—such as clinical interviews, self-report questionnaires, and counselling sessions—remain essential but are often limited by stigma, underreporting, and lack of timely access to professionals. Many adolescents hesitate to seek help due to fear of judgment or lack of awareness, instead turning to digital spaces where their distress may be more openly expressed. This shift highlights the potential of technology-driven solutions to complement conventional methods.

In recent years, Machine Learning (ML) and Artificial Intelligence (AI) have emerged as powerful approaches for analysing the vast volumes of unstructured data—text, audio, and video—produced on social media and digital communication platforms. By applying Natural Language Processing (NLP), computer vision, and speech analysis, ML models are able to detect subtle linguistic markers, facial expressions, and vocal patterns that may indicate depression or suicidal ideation (Chancellor and De Choudhury, 2020). Pestian et al. (2012), for example, demonstrated that ML algorithms could outperform human experts in differentiating authentic suicide notes from fabricated ones. More recently, a systematic review and meta-analysis published in the *Journal of Medical Internet Research* (2025) reported that ML models achieved strong predictive accuracy, with AUC values ranging from 0.77 to 0.84, in identifying suicidal ideation and attempts among adolescents, based on data from over 1.4 million participants across 42 studies. Similarly, a review in *BMC Medical Informatics and Decision Making* (2024) highlighted that ensemble algorithms such as Random Forest and XGBoost consistently outperformed other methods, achieving AUC scores as high as 0.97 in suicide risk prediction.

The COVID-19 pandemic further amplified the urgency of such approaches. Lockdowns, social isolation, and academic disruptions significantly worsened mental health outcomes for students and teenagers worldwide. During this period, online platforms became not only a means of social interaction but also a critical window into the psychological states of young people. This reinforced the importance of developing real-time, scalable, and ethically responsible ML-based systems that can detect early warning signs and trigger timely interventions.

Despite promising advances, significant research gaps remain. Existing models often struggle with data scarcity, cultural biases, and ethical concerns related to privacy and consent. Moreover, while text-based detection has been widely studied, the integration of multimodal data—combining social media posts with facial expressions and speech cues during video calls—is still underexplored. A systematic review in MDPI Sensors (2024) and a survey in ACM Computing (2025) both emphasize that multimodal ML approaches (text + audio + video) consistently outperform single-modality models, yet their application in adolescent suicide prevention remains limited. Addressing these gaps is crucial for creating robust, inclusive, and context-sensitive solutions that can be applied across diverse populations.

This study aims to examine the role of machine learning in detecting suicidal tendencies and depression among students and adolescents, with a particular focus on analysing both social media posts and video call interactions. By situating the research within the broader global discourse on suicide prevention, while grounding it in the specific digital behaviours of young people, the study seeks to contribute to both academic scholarship and practical applications in mental health care.

In recent years, several studies have explored the impact of social media on suicidal ideation. Choudhury et al. developed a statistical model based on score-matching techniques to identify transitions from general mental health discussions to suicidal thoughts. Their findings suggest that this shift can be characterised by three psychological stages: *thinking*, marked by anxiety, despair, and misery; *ambivalence*, associated with reduced social connectedness and low self-esteem; and *decision-making*, often accompanied by aggression and the formulation of a suicide plan (Aldhyani et al., 2022).

Globally, suicide remains a critical public health issue, with nearly 800,000 deaths reported annually. The worldwide suicide rate is estimated at 10.5 per 100,000 population, making it the second leading cause of death among young people. Alarmingly, approximately 79 per cent of suicides occur in low- and middle-income countries, where resources for detection, prevention, and treatment are often limited (World Health Organization, 2018).

Suicidal ideation refers to the tendency to contemplate taking one's own life, ranging from transient feelings of sadness to persistent preoccupation with self-destruction and, in some cases, the formulation of a suicide plan. Within research communities, individuals are often categorised as *suicide ideators* (planners) or *suicide attempters* (completers), and the relationship between these groups has been the subject of considerable debate. Evidence suggests that while many individuals experience suicidal thoughts, only a minority progress to actual attempts. Klonsky et al. (2016), for example, argue that commonly cited risk factors such as hopelessness, despair, and frustration are more predictive of suicidal ideation than of the transition from ideation to attempt. Conversely, Pompili et al. (2011) highlight that ideators and attempters may share several overlapping risk factors, complicating distinctions between the two groups. Recognising this, the World Health Organization (2018) has encouraged member states to adopt early identification of suicidal thoughts as a cornerstone of national suicide prevention strategies, with the aim of reducing global suicide rates.

In recent years, social media has emerged as a powerful lens into the mental health and wellbeing of its predominantly young user base. By enabling anonymous participation in online communities, these platforms provide opportunities for open discussion of stigmatised topics such as depression and suicidality (Nayini et al., 2023). Written expressions of suicidality are particularly concerning, as suicide notes are left in approximately half of completed suicides and in over 20 per cent of attempts. Unlike retrospective offline texts, social media posts—such as blogs, tweets, and forum messages—are often contemporaneous and well-preserved, offering valuable data for analysis (Choudhury et al., 2016).

The study of online discourse has therefore become an important area within computational linguistics, providing a fertile environment for the development of innovative technologies aimed at suicide detection and prevention. For instance, Kumar et al. (2015) examined the posting behaviours of Reddit Suicide Watch users in the context of celebrity suicides, proposing strategies to mitigate the risk of contagion effects. Similarly, Choudhury et al. (2016) employed propensity score matching to identify linguistic markers signalling a shift from general mental health discussions to suicidal ideation. More recently, Ji et al. (2022)

introduced advanced data-protection and optimisation techniques to enhance the early detection of suicidal thoughts.

Beyond traditional text classification methods, deep learning approaches have achieved significant progress in natural language processing (NLP) and pattern recognition. Neural networks based on dense vector representations, such as word embeddings, have consistently outperformed conventional machine learning models that rely on handcrafted features. Architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) models have demonstrated particular promise in assessing suicide risk from unstructured social media data (Tadesse et al., 2019).

1.3 Problem Statement

Although suicide is increasingly recognised as a leading cause of death among adolescents and young adults, prevention strategies remain insufficient. Conventional approaches such as clinical interviews, self-report surveys, and counselling interventions are often constrained by stigma, underreporting, and limited accessibility (Patel et al., 2018). Many students and teenagers do not seek professional support, yet they frequently disclose psychological distress through their digital footprints on social media platforms and during online interactions. This duality presents both a challenge and an opportunity: while vast amounts of behavioural data are available, existing systems for detecting suicidal tendencies are fragmented, reactive, and frequently fail to provide timely intervention.

Recent advances in Machine Learning (ML) and Artificial Intelligence (AI) have demonstrated considerable potential in identifying patterns of suicidal ideation. Evidence suggests that ML models can achieve high predictive accuracy in detecting suicidal content from text-based data (*Journal of Medical Internet Research*, 2025; *BMC Medical Informatics and Decision Making*, 2024). However, the majority of existing research has focused narrowly on textual analysis of social media posts, leaving significant gaps in the integration of multimodal data sources such as facial expressions, vocal tone, and behavioural cues during video calls. This limitation is critical, as adolescents often communicate distress nonverbally or through subtle signals that text-only models cannot capture (Chancellor and De Choudhury, 2020).

The urgency of this issue is particularly evident in India, where student suicides have reached alarming levels. The National Crime Records Bureau (2024) reported more than 13,000 student suicides in a single year, with rates rising faster than the national average. A longitudinal analysis published in *Child and Adolescent Psychiatry and Mental Health* (2024) further projected a continuing upward trajectory in youth suicides over the next decade. Despite this escalating crisis, there remains limited research on how AI-driven, multimodal detection systems can be designed and implemented in culturally sensitive and ethically responsible ways to support early intervention.

Therefore, the central problem this study addresses is the lack of comprehensive, multimodal machine learning frameworks capable of detecting suicidal tendencies and depression among students and teenagers in real time, across both social media posts and video call interactions. Without such systems, opportunities for timely intervention are missed, perpetuating the cycle of under-detection and preventable loss of life.

Categories of Suicides

One type of suicide, called **egoistic suicide**, occurs when a person believes they have been socially isolated and, as a result, feels unappreciated in their society. This type of suicide is typically associated with people who are self-absorbed and lacking in altruism, who have been cut off from the mainstream of society.

The second kind, **altruistic suicide**, involves a greater degree of social cohesion amongst those who are tied to the proof on some level, as shown, for example, in Dannie's warriors or the Sati system.

Suicide attempts in the wake of financial hardships or unexpected good fortune (such as winning the lottery) are examples of anomic suicide. The event arises unexpectedly as a result of a certain activity.

For example, in a fatalistic society where everyone is forced to work for someone else, those who are unable to have children could take their own lives because they feel they have no other choice. Some indicators however, could be:

- Numerous warning indicators and indicators of suicide
- Abrupt shift in eating and sleeping patterns
- Isolation from close personal and social relationships and from daily routines.
- Acts of defiance, violence, or flight from authority.
- Substance abuse and alcoholism
- They have an abnormal lack of care in how they present themselves.
- Transformative shifts in character traits.
- Ongoing decline in productivity and quality due to boredom and trouble focusing.
- Complaints about the bodily manifestations of different emotions, such as a headache, stomachache, or extreme weariness.
- A sudden lack of enthusiasm for things that used to bring you joy.

- Not motivated by the prospect of future acclaim or material gain.

Pre-Suicidal Factors

Recent research has increasingly examined the psychological impacts of social media platforms such as Facebook, Instagram, and Twitter. According to a review by *BBC Future* (2019), approximately 40 per cent of the global population—equivalent to around three billion individuals—are active users of online social networks. On average, people spend nearly two hours per day engaging with these platforms, posting updates, responding to friends, and sharing personal content. The scale of activity is immense, with users collectively generating around 500,000 tweets and Snapchat images every minute. Given this level of pervasiveness, understanding the effects of social media use on psychological wellbeing has become a critical area of inquiry.

STRESS

People vent their frustrations about anything from subpar customer service to societal issues on social media. While cathartic, this results in a never-ending parade of complaints on our news feed. Recent studies have aimed to determine whether or not our online social lives relieve stress or amplify bad feelings.

ANXIETY

The team also aimed to understand how interconnections with others contribute to generalized stress. People who use seven or more networks have a general anxiety level three times greater than users who use two social media platforms or fewer, according to a research published in the journal Computers and Human Behavior (Computers and Human Behaviour).

DEPRESSION

Previous research has linked social media usage to feelings of despair, but recent studies have revealed the reverse to be true. For example, research involving over 700 students shows that low-quality online communication is linked to depressive symptoms including low mood, feelings of inadequacy, and hopelessness. Depressive symptoms were more severe in individuals whose experiences with virtual contact were mostly unfavorable. In 2016, researchers surveyed 1,700 adults and found that heavy social media users were three times as likely to suffer from despair and anxiety. Researchers highlighted many factors, but cyberbullying and a skewed perception of others' lives stood out as particularly significant. (Bansal et al., 2024)

Researchers have increasingly explored how social media can be utilised to identify depressive symptoms and encourage timely treatment-seeking behaviours. One notable study commissioned by Microsoft analysed the accounts of 476 Twitter users to detect indicators of depression, including tweet sentiment, emotional intensity, patterns of interaction with other users, and posting frequency. Based on these digital markers, the researchers developed a diagnostic questionnaire that was able to predict the likelihood of depression with approximately 70 per cent accuracy, even in cases where individuals displayed no overt clinical symptoms (De Choudhury et al., 2013).

LONELINESS

Evidence suggests that excessive use of social networking platforms may be associated with heightened feelings of social isolation. A large-scale study published in the *American Journal of Preventive Medicine* involving 7,000 individuals aged 19–32 found that participants who spent more time on social media were nearly twice as likely to report perceived social isolation compared with those who used such platforms less frequently (Primack et al., 2017). The findings highlight the paradox that, despite offering opportunities for connection, heavy engagement with social media may hinder the development of meaningful interpersonal relationships.

Researchers found that individuals who spent too much time on social networks were more likely to report feeling lonely than those who spent less time online. When we look at the lives of our friends and acquaintances through rose-colored glasses, we may end up feeling envious and mistakenly believing that they are happier and more successful than we are. Isolation increases when people think this way. (Bonsaken et al., 2023)

How Are Social Networks And Dysmorphophobia Related?

Body dysmorphic disorder (BDD) is characterized by extreme self-criticism and even hate of one's physical appearance; symptoms include an obsession with finding flaws in one's physical appearance (often via the use of social media), anorexia, and other eating disorders. The main argument is that people's perceptions of their own bodies become "somewhat different" as a result of constantly being exposed to distorted body images on Instagram.

More and more medical professionals, including plastic surgeons, are claiming that social media encourages more surgeries. They now see a new kind of dysmorphophobia in which individuals present not with celebrity photographs but with snapchat filters. Meanwhile, body positivity accomplishes its goal by providing a forum for people whose physical attributes have not been reflected in the media to share their stories with the world. These individuals go on to star in ad campaigns, give speeches, and serve as role models, proving that their bodies are deserving of respect.

Researchers have found the following patterns of social network behavior among those who exhibit depressive symptoms. Several behavioural symptoms have been identified in relation to problematic social media use and its association with depressive tendencies among students.

- **Social comparison:** Many individuals frequently evaluate themselves against other users who appear to lead more desirable lives. Platforms such as Instagram, where users often share selectively positive images, have been identified as hotspots for this phenomenon (Fardouly et al., 2015).

- **Excessive use and dependency:** Social networking can reach levels resembling behavioural addiction. Survey data indicate that students who acknowledge attempts to reduce their usage, but report negative impacts on academic or professional performance, demonstrate patterns consistent with problematic use (Andreassen et al., 2017).

- **Anxiety linked to online feedback:** Students often experience heightened anxiety when their online posts, particularly photographs, fail to receive positive attention. This unease has, in some cases, been associated with frustration and even aggressive behaviour (Marino et al., 2018).

- **Reduced social sharing:** Depressed individuals are less likely to post group photographs online, reflecting tendencies toward social withdrawal and reduced participation in collective activities (Lin et al., 2016).

Applying AI in suicidal ideation detection and suicide prevention

Suicide remains one of the most pressing public health concerns, ranking as the third leading cause of death among individuals under the age of 25 (World Health Organization, 2018). In response, researchers have increasingly explored the application of Artificial Intelligence (AI) and Machine Learning (ML) in suicide prevention. Early attempts date back to 2007, when investigators analysed data from MySpace users to identify individuals at risk of suicide or recent attempts, demonstrating modest success (Ohman and Watson, 2019).

AI-driven approaches have since advanced considerably. For example, linguistic markers such as vowel length during speech have been shown to indicate suicidal inclination, while ML algorithms have outperformed clinicians in distinguishing genuine suicide notes from fabricated ones, achieving 78 per cent accuracy compared with 63 per cent for human experts (Pestian et al., 2010). Despite decades of research, traditional risk factors such as depression, stress, and substance use have proven insufficient for accurate prediction, often performing no better than chance. This limitation has been attributed to overly simplistic models that fail to account for the complex interplay of multiple risk factors.

Recent breakthroughs have demonstrated the potential of neuroimaging combined with ML. Using functional magnetic resonance imaging (fMRI), researchers in the United States applied ML algorithms to brain activity data from individuals with suicidal ideation, achieving a detection accuracy of 91 per cent (*Nature Human Behaviour*; Wein, 2017). Further studies have explored how brain representations of suicide-related and emotional concepts differ between healthy individuals and those with suicidal tendencies. For instance, the parahippocampal gyrus, associated with processing real-world scenes, has been implicated in concept comprehension such as “home.” These findings suggest that variations in neural activation patterns may be predictive of suicidal behaviour (Aminoff et al., 2013).

In one study, a naïve Bayes classifier was applied to voxel-level fMRI data, identifying significant differences in activation between 17 healthy participants and 17 individuals with suicidal ideation. Six concepts—death, cruelty, problems, carelessness, goodness, and praise—were found to elicit the most distinct neural responses between the groups (Aminoff et al., 2013).

The broader context underscores the urgency of these advances. Depression and anxiety are rising globally, with particularly severe impacts in industrialised nations. The World Health Organization (2018) estimates that nearly 800,000 people die by suicide each year, making it one of the leading causes of death among those aged 15–29. Contributing factors include work-related stress, loneliness, schizophrenia, social isolation, and other adverse life events. Suicidal ideation—defined as thoughts of taking one’s own life—is a common symptom of untreated mental illness and is especially prevalent among young people. Increasingly, these individuals disclose suicidal thoughts through digital channels such as social media, blogs, text messages, and online forums, rather than in face-to-face conversations (Nayini et al., 2023).

Extensive analysis of this data reveals that the use of anonymous online platforms to vent frustration will continue it's rise in the coming years. The question of how much suicide danger is really caused by these kinds of online utterances becomes even more murky. Although researchers have attempted to quantify this threat and examine the association between the two, their studies have focused on one or two aspects of suicidal thoughts rather than the whole spectrum.

Digital data such as text messages, social media posts, and blog entries have become valuable resources in the study and evaluation of suicidal ideation. The relative anonymity of online platforms often enables individuals to disclose thoughts and emotions they might otherwise withhold in face-to-face interactions. This user-generated content can therefore support the early detection of suicidal thoughts and facilitate timely therapeutic interventions. Increasingly, digital platforms have begun to monitor such expressions and mine social information to advance suicide prevention research.

However, these same platforms have also given rise to concerning social phenomena, including online communities that promote self-harm or encourage copycat suicides. A notable example is the "Blue Whale Challenge," a social media trend reported in 2016 that allegedly encouraged participants to complete a series of tasks, some involving self-harm, culminating in suicide (Khasawneh et al., 2020). Such cases underscore the dual role of digital environments as both potential tools for prevention and spaces where harmful behaviours may be amplified.

Given that suicide claims the lives of thousands of individuals each year, it remains a critical global public health issue. Recognising early warning signs of suicidality and intervening promptly is essential, as timely action can help prevent suicide attempts and save lives (World Health Organization, 2018).

Current State of the Art

In response to the growing prevalence of suicide-related content online, major social media platforms have begun implementing tools to identify and respond to signs of suicidal ideation. For example, in 2016 Facebook introduced a feature that enables users to flag posts they perceive as "suicidal or self-harming." Once flagged, the platform provides several options, including notifying Facebook for review, offering contact details of suicide prevention organisations, and allowing users to send supportive messages to the individual in distress. In cases of imminent danger, Facebook advises contacting local emergency services immediately (Facebook, 2016).

While such initiatives represent important steps, their effectiveness is limited in developing regions where technological adoption by users, emergency services, and support agencies remains relatively low. This creates challenges in ensuring timely alerts and interventions.

Moreover, reliance on manual reporting by other users is not scalable for platforms serving billions of accounts worldwide. To address these limitations, researchers and industry leaders have advocated for automated, AI-driven approaches capable of detecting suicidal content in real time by analysing posts and associated comments (Chancellor and De Choudhury, 2020).

Research Problem

Early identification of suicide thoughts in depressed people may provide appropriate medical care and support, which is often life-saving. Recent NLP research focuses on identifying if a person is clinically healthy or suicidal from a given text. However, despite being a significant therapeutic concern, there have been little efforts to distinguish between depression and suicidal thoughts. Web query data has become a possible substitute due to the limited availability of EHR data, suicide notes, or other such confirmed sources. Online sources like Reddit, which provide anonymity and encourage open discussion of symptoms, are credible sources even in a therapeutic environment. The inherent noise in web-scraped labels, which calls for a noise-removal technique, is the cause of these online datasets' inferior performance, however.

As a result, we suggest SDCNL (Suicide vs Depression Classification), a deep learning-based classification strategy that distinguishes between depression and suicide. We use online material from Reddit to train our system, and we offer a unique unsupervised label correction strategy that, in contrast to other work, does not need knowledge of the noise distribution in advance. Our exhaustive testing of several deep word embedding models and classifiers demonstrates the method's potent performance in a novel, difficult classification application. We publish our source code and data at https://github.com/ayaanzhaque/SDCNL.

1.3 Purpose of Research

The overarching objective of this study is to investigate the role of machine learning in detecting suicidal tendencies and depression among adolescents and students. Specifically, the research focuses on analysing both social media posts and video call interactions as key sources of behavioural and emotional data. The study seeks to design and evaluate a multimodal framework that integrates textual, visual, and auditory inputs, thereby enabling more accurate and timely identification of psychological distress. By combining these modalities, the framework aims to advance current approaches to suicide prevention and contribute to the development of proactive, technology-driven mental health interventions.

1.4 Research Objectives

To achieve its overarching aim, this study pursues several specific objectives:

1. To critically review existing literature on suicide prevention, adolescent mental health, and the application of machine learning in risk detection.
2. To analyse the limitations of current text-based detection models and identify opportunities for multimodal approaches.
3. To develop a conceptual framework that integrates social media text, facial expressions, and vocal features into a unified ML-based detection system.
4. To evaluate the performance of selected machine learning algorithms (e.g., Random Forest, XGBoost, CNN, LSTM) in detecting suicidal tendencies across different data modalities.
5. To assess the ethical, cultural, and practical considerations of applying ML-based detection systems in adolescent populations, with particular reference to the Indian context.
6. To propose recommendations for policymakers, educators, and mental health practitioners on the responsible use of AI in suicide prevention.

The primary goal is to demonstrate how machine learning can be applied to the automated detection of suicidal messages. This study focuses particularly on Twitter and Reddit, where users frequently disclose distress through short posts and comments. The aim is to design a system capable of categorising social media messages according to indicators of suicidality, thereby identifying individuals who may be struggling with suicidal thoughts. While conventional suicide prevention strategies rely heavily on clinical approaches such as self-reports and in-person interviews with therapists or social workers, these methods are often limited by stigma, underreporting, and accessibility barriers.

Machine learning offers a novel approach to early identification by enabling the automated recognition of suicidal content in online discussions. Numerous factors—including stress, low self-esteem, and mental distress—contribute to suicidal ideation, and adolescents are particularly vulnerable to these risks. By applying feature engineering, sentiment analysis, and deep learning techniques, ML models can detect patterns in social media data that may indicate suicidal intent (Aldhyani et al., 2022). Neural networks, particularly CNNs and LSTMs, have shown promise in learning rich representations of language associated with suicidality.

Beyond text, suicide prevention research has also explored mobile and digital interventions. Examples include the Black Dog Institute's *iBobbly* app, the *Virtual Hope Box*, and *Suicide Safe*, which provide therapeutic support via mobile platforms. Social networking-based tools such as *Samaritans Radar* (a Twitter plug-in later discontinued due to privacy concerns) and *Woebot* (a Facebook chatbot using cognitive behavioural therapy and NLP) illustrate both the potential and the ethical challenges of AI-driven interventions.

Ethical concerns remain central to the deployment of such systems. Linthicum et al. (2020) identified three key challenges: algorithmic bias, the prediction of suicide timing, and the ethical/legal implications of false positives and false negatives. These dilemmas highlight the difficulty of balancing competing interests, values, and risks in AI-based suicide prevention. While AI has been applied to numerous societal challenges, its use in detecting suicidal ideation requires particular sensitivity to privacy, cultural context, and the potential consequences of misclassification.

Despite progress, standards for training and evaluating suicide ideation detection models remain limited. AI systems may detect statistical signals but cannot “read minds,” and many neural models lack interpretability. This study therefore examines multiple approaches to suicidal thought detection, with a focus on domain-specific applications that carry significant societal implications.

The urgency of this research is underscored by alarming statistics. According to the National Crime Records Bureau (S. Biswas, 2020), in India alone, 26,589 daily wage earners, 12,175 self-employed individuals, and 10,687 unemployed men died by suicide in 2018. Among women, 22,937 homemakers, 4,790 students, and 3,535 wage employees were recorded as suicide cases. Even among transgender individuals, suicides were reported across occupational categories. The suicide rate in India is estimated at 16.5 per 100,000 (Colic, 2018). Globally, the World Health Organization (2018) reports that more than 800,000 people die by suicide each year—equivalent to one death every 40 seconds—yet only 30 per cent disclose their intentions beforehand.

Given the scale of the problem, there is a pressing need for strategies that can detect suicidality early and intervene effectively. Linguistic analysis, including the study of speech production and grammatical structures, has been proposed as a diagnostic tool for uncovering subconscious indicators of vulnerability. With the rapid growth of social networks, the volume and velocity of user-generated content provide both opportunities and challenges for suicide detection. Big data analytics now makes it possible to organise seemingly random information into meaningful patterns, enabling supervised and unsupervised ML algorithms to be applied at scale.

This study evaluates the relative efficacy of such algorithms in detecting suicidal ideation on VKontakte, one of the most widely used social networks among Kazakh youth. By comparing classification accuracy across algorithms, the research seeks to identify the most effective models for distinguishing suicidal from non-suicidal posts. The ultimate aim is to provide evidence-based guidance on selecting appropriate ML techniques for suicide detection in social media contexts.

Finally, the study situates suicidal ideation within its broader conceptual framework. Suicidal thoughts, encompassing both depressive states and specific plans, remain a contested area of research. Klonsky et al. (2016) argue that common risk factors such as depression and hopelessness predict ideation but not attempts, whereas Pompili et al. (2011) suggest that ideators and attempters share many overlapping traits. In line with the World Health Organization's global target of reducing suicide rates by 10 per cent by 2020, early detection tools have been prioritised internationally. Advances in sentiment analysis and NLP have enabled the automatic extraction of emotional signals from social media data, offering new opportunities for prevention (Aldhyani et al., 2022).

1.5 Significance of the Study

1. **Introduction: Problem Statement and Research Goals**

- **Problem:** Suicidal ideation expressed on social media platforms such as Reddit and X (formerly Twitter) has emerged as a pressing public health concern. While early detection and timely intervention are critical for suicide prevention, conventional approaches—including clinical interviews, surveys, and manual monitoring—are often inadequate for analysing the vast and complex volumes of linguistic data generated online. This gap underscores the need for advanced computational methods capable of processing unstructured digital content at scale and identifying subtle indicators of psychological distress.
- **Research Objective:** The primary objective of this study is to design, develop, and evaluate deep learning models for the accurate and efficient detection of suicidal ideation within social media text. By leveraging state-of-the-art natural language processing (NLP) techniques, the study seeks to contribute to both academic research and practical applications in suicide prevention.

- **Specific Goals:**
  To achieve this objective, the study will pursue the following goals:
  - To explore and compare the performance of Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and a hybrid LSTM-CNN architecture for multi-class classification of suicidal ideation severity.
  - To investigate whether the hybrid LSTM-CNN model demonstrates superior performance compared with individual LSTM and CNN models, as well as traditional machine learning baselines.

- To assess the generalisability of the developed models across multiple social media platforms.
- To design a flexible detection framework that can be adapted for use on diverse online forums, blogs, and digital communication platforms.

2. Literature Review: Context and Existing Approaches

- **Social Media and Suicide Ideation:**
  - Research has increasingly demonstrated that online linguistic and social patterns can serve as indicators of suicide risk. Coppersmith et al. (2016) identified distinctive linguistic markers, such as heightened expressions of sadness, anger, and hopelessness, that are strongly associated with suicidal ideation. Beyond individual language use, the dynamics of online social networks also play a critical role.
  - Studies by Hsiung (2007), Jashinsky et al. (2014), and Colombo et al. (2016) emphasise how patterns of connectivity and peer interaction within digital communities can amplify suicidal thoughts and behaviours. External events further influence these dynamics: the "Werther effect," whereby media coverage of celebrity suicides triggers increases in suicidal behaviour, has been observed in online contexts.
  - Kumar et al. (2015) and Ueda et al. (2017) provide evidence that celebrity suicides can significantly shape online discourse and, in some cases, contribute to subsequent increases in suicide rates. Collectively, these studies highlight the complex interplay between individual expression, social connectivity, and external cultural events in shaping suicide risk within digital environments.

- **Machine Learning for Suicide Ideation Detection:**
  - Traditional machine learning approaches have been widely applied to the detection of suicidal ideation, though with varying levels of success. Algorithms such as Support Vector Machines (SVMs), logistic regression, and Random Forests have been employed to classify at-risk individuals based on linguistic and behavioural features extracted from online data. Several studies have explored the effectiveness of feature engineering and classification techniques in this domain. For example, Braithwaite et al. (2016), Sueki et al. (2014), O'Dea et al. (2015), and Wood et al. (2001) investigated the predictive potential of structured features derived from text and survey data.

    - In parallel, information retrieval methods such as term frequency–inverse document frequency (TF-IDF) have been applied to represent textual content, as demonstrated in the work of Okhapkina et al. (2019). Collectively, these studies highlight both the promise and the limitations of conventional ML approaches, particularly their reliance on handcrafted features and their difficulty in capturing the nuanced semantics of suicidal expression.like Okhapkina et al.

- **Deep Learning Advancements:**
    - Recurrent Neural Networks (RNNs), particularly LSTMs, excel at capturing sequential dependencies in text.
    - Convolutional Neural Networks (CNNs) are effective in extracting local n-gram features.
    - Hybrid models combining LSTM and CNN architectures have shown promise in various NLP tasks.

3. Methodology: Proposed Model and Evaluation

- **Dataset:**
    - The study will utilize datasets from Reddit (e.g., SuicideWatch forum) and X, known for their high volume of user-generated text.
    - Data preprocessing will involve cleaning, normalization, and tokenization to prepare the text for model training.
- **Model Architecture:**
    - As part of the experimental design, individual deep learning models will first be developed and trained using Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). Building on these baseline models, a hybrid LSTM-CNN architecture will then be constructed to leverage the complementary strengths of both approaches.
    - In this framework, the CNN component will be employed to extract localised features from the input text, while the LSTM component will capture the sequential dependencies and contextual relationships among those features. To ensure optimal performance, systematic hyperparameter tuning will be conducted, including adjustments to learning rates, batch sizes, dropout rates, and the number of hidden units. This process will allow for the identification of the most effective model configuration for detecting suicidal ideation in social media data.

- **Evaluation Metrics:**

    - The performance of the developed models will be evaluated using standard classification metrics, including precision, recall, F1-score, and overall accuracy. These measures will provide a comprehensive assessment of the models' ability to correctly identify suicidal ideation while minimising false positives and false negatives.
    - To ensure the robustness and reliability of the results, cross-validation techniques will be employed. In particular, k-fold cross-validation will be used to reduce the risk of overfitting and to provide a more generalisable estimate of model performance across different subsets of the data. This evaluation strategy will enable a rigorous comparison of the baseline and hybrid models, ensuring that the findings are both statistically sound and practically meaningful.
    - To demonstrate the advantages of deep learning models, their outcomes will be contrasted with those of conventional machine learning models.

- **Multi-Class Classification:**
    - The models will be designed to categorize suicide ideation into multiple levels of severity, enabling a more sophisticated comprehension of risk.

4. Expected Contributions and Implications

- **Improved Detection Accuracy** The proposed hybrid LSTM-CNN model is expected to achieve higher levels of accuracy in detecting suicidal ideation when compared with individual deep learning models and conventional machine learning approaches. By combining the strengths of both architectures, the model aims to capture nuanced linguistic and contextual features that are often overlooked by traditional methods.

- **Enhanced Understanding of Linguistic Patterns** Beyond predictive performance, the study will contribute to a deeper understanding of the linguistic markers associated with suicidal ideation in social media discourse. Insights into these patterns may inform both computational modelling and clinical perspectives on how distress is expressed online.

- **Development of a Practical Tool** The models developed in this research have the potential to be integrated into online platforms as practical tools for early detection and intervention. Such systems could support mental health practitioners, educators, and policymakers in providing timely assistance to individuals at risk.

- **Generalisability Across Platforms** A key goal of the study is to demonstrate the adaptability of the proposed framework across multiple social media platforms.

By testing the model on diverse datasets, the research seeks to establish its robustness and applicability in varied digital environments.

- **Ethical and Privacy Considerations** The study will adhere strictly to ethical guidelines, with particular emphasis on safeguarding user privacy and ensuring responsible use of sensitive data. Ethical considerations will be central to the design, evaluation, and potential deployment of the proposed system, acknowledging the societal implications of AI-driven suicide detection.

5. Future Work

- Investigating how contextual data, including user demographics, can be integrated and social network interactions, to further improve model performance.
- Investigating the use of attention mechanisms and transformer-based models for suicide ideation detection.
- Developing real-time detection systems for online platforms.
- Examine how various word embedding types affect the model’s functionality.

1.6 Research Purpose and Questions

The purpose of this research is to examine methods through which user-generated web material may be used to better comprehend suicidal thoughts, allowing for earlier suicide detection via supervised learning.

When combined with the topic description, a user's preferred language can reveal a wealth of information that can help an early warning system identify suicidal tendencies (Dey, R. K., 2020). This is significant because suicidal individuals frequently exhibit anxiety and hopelessness. Friends and family are likely to be involved, and as a result, the conversations will likely center on social and personal issues; these can be identified using linguistic, syntactic, and statistical embedded-word sets, and six classifiers, including the four supervised classifiers traditionally used and neural networks, can be compared to one another (Shaoxiong Ji Meng, 2019).

The experimental research would prove the approach's viability and usefulness, establishing a standard for measuring suicide thoughts on online platforms including Twitter, Reddit, and Suicide Watch (Gradus, 2020).

Adolescents have a high rate of nonfatal suicide attempts, making it difficult to clinically provide prediction of risk in practice. Screening related to time consumption that includes implementation scale is necessary so that trained counselors can offer crisis-tailored management. Analysis of the user's text psycholinguistics using Word Count and "Simplified Chinese Linguistic Inquiry" would take place in the following one month after consultation (Hosseinifard, 2013).

For the random strategy of forest to be accepted by the community of Machine Learning and applied simply along with its robustness (Sarddar, D.,2020)(Bose, R.,2020), preprocessing the data would be performed using the technology Python, which would analyze the statistics and perform them in R.

AUC, coupled with the metrics recall and accuracy, would be used to assess the model's efficacy. The study's accuracy in making predictions may be gauged by looking at calibration plots and weighing Brier scores, which can vary from 0 to 1 (Sharmistha Dey, 2020; Biswas, S., 2020).

Merchant (2018) conducted a study examining the relationship between internet use and self-harm, analysing publications from the preceding four years. The review drew upon multiple studies involving individuals with experiences of self-harm or suicidal ideation. While the findings indicated a correlation between online activity and suicidal thoughts, the study did not identify which specific platforms were most responsible for this association.

Building on these gaps, Ruder et al. (2011) investigated the role of Facebook in suicidal behaviour. Their research focused on the phenomenon of “Facebook suicide notes,” exploring their causes, potential for copycat behaviour, and implications for prevention strategies. Although several instances of suicide notes posted on Facebook were identified, the study found no conclusive evidence of copycat suicides, making it difficult to determine which cases required urgent intervention.

Researchers Han et al., (2017) want to know how frequently people have suicide thoughts. A 12-month poll revealed that 39% of respondents kept their suicide thoughts to themselves, 47% shared their struggles on social media, and 42% saw a doctor. The 39% of respondents who chose to remain anonymous have not been helped by this study.

Researchers Luoma et al., (2001) looked at the number of times doctors and other medical staff interact with patients in the days and weeks before they take their own lives. According to the findings, over half of suicide victims visit a doctor within the month before they pass away. However, it did not specify how effective these doctors and nurses are in preventing suicide.

Researchers Phaltankar et al., (2022) investigated the many barriers to conducting suicide risk assessments and finding people who have suicidal thoughts. According to the research, increased teen suicide rates are attributable to inadequate screening for suicidality and inadequate training of first responders, despite the availability of medical care providers.

The role of various technologies in the identification and prevention of suicidal thoughts was investigated in a separate survey by Franco-Martín et al., (2020). While the findings indicated that different platforms, such as the online, mobile, primary care, and social networks, all had a role to play in improving suicide ideation detection, they also showed that doing so together would be most effective.

Hawton et al., (2018) presented a study demonstrating the prevalence of overt suicide attempts among survey respondents.

Our last poll was based on findings from a study by Renjith et al., (2022) This study employs a real-time database and deep learning methods to create an emotion-detection LSTM-CNN model for identifying suicidal thoughts.

Suicidal behavior in adolescents

Beaven-Ciapara et al. (2018) examined risk factors for suicidal ideation and behaviour among adolescents in Guaymas, Sonora, Mexico. Their study, conducted with 120 middle and high school students (41% male and 59% female) using SPSS 21.0, found that suicidal behaviour was associated with psychosocial factors in 37% of cases, with a higher prevalence among females. The findings suggested that dysfunctional family environments were a key contributor, generating despair and low self-esteem.

Complementing this, Mosquera (2016) provided a non-systematic review of child and adolescent suicidal behaviour, identifying male gender, prior suicide attempts, social isolation, and emotional conflict as significant risk factors. High comorbidity with mood disorders, mania, and psychosis was also observed. Among the most effective interventions highlighted were cognitive behavioural therapy (CBT) and dialectical behaviour therapy (DBT).

Carballo-Belloso and Gómez-Peálver (2017) further emphasised the role of psychosocial stressors, establishing a strong causal link between childhood bullying experiences, individual vulnerability factors, and the subsequent development of suicidal thoughts or self-inflicted behaviours. This underscores the importance of detecting and addressing modifiable risk factors early.

Parallel to psychosocial research, machine learning methods have been increasingly applied to suicide detection. O’Dea et al. (2015) identified 14,701 suicide-related tweets with an accuracy of 80%, while Wang et al. (2019a) conducted a similar study on Chinese social media. Benton et al. (2017b) employed a feed-forward neural network trained on character n-gram features within a multi-task learning (MTL) framework to predict suicidal ideation and abnormal mental health. Johnson Vioulès et al. (2018) proposed a natural language processing (NLP) system that rapidly identified suicide-related posts on Twitter using a lexical ensemble approach. Comparable studies by Cheng et al. (2017) and Huang et al. (2014) also demonstrated strong associations between linguistic features and suicidal expression.

At a population level, Jashinsky et al. (2014) found that the frequency of suicide-related phrases in tweets correlated with age-adjusted suicide rates across U.S. states. Other predictors, such as prior suicide attempts (Rakesh, 2017) and self-harm (Robinson et al., 2015), have been consistently identified as substantial risk factors for completed suicide.

De Choudhury et al. (2016) analysed Reddit users who shifted their focus from general mental health discussions to suicidal ideation, applying cognitive psychological integrative models of suicide (reasoning, ambivalence, and decision-making) to identify early warning signs of this transition. Similarly, Roy et al. (2020) developed a model that incorporated ten psychological factors—including stress, loneliness, burdensomeness, hopelessness, depression, anxiety, and insomnia—alongside sentiment polarity. These features were used to train a Random Forest classifier, achieving an AUC score of 0.88, demonstrating the potential of combining psychological metrics with machine learning for suicide risk prediction.

Suicide analysis with Artificial Intelligence

Ramírez-López et al. (2021) utilised a SQL Server database of 911 emergency service records to anticipate potential suicide cases in Aguascalientes, Mexico. Applying Weka's Empirical Bayesian Kriging (EBK) method, they conducted a geospatial analysis to identify high-risk areas. A Bayesian classifier implemented in MATLAB achieved a prediction accuracy of 99.22%, providing a visual representation of probability regions where suicides were more or less likely to occur.

Given that military personnel experience higher rates of psychological stress and suicide attempts than the general population, Gen-Min et al. (2020) investigated suicidal ideation in this demographic. Using responses from 3,546 participants, they applied a range of machine learning techniques—including logistic regression, decision trees, random forests, regression trees, support vector machines, and multilayer perceptrons—to analyse five psychopathological domains (BSRS-5): anxiety, depression, hostility, interpersonal sensitivity, and insomnia. Their models achieved classification accuracies exceeding 98%.

Pérez-Martínez et al. (2020) examined discourse on Twitter relating to school bullying, rape, and suicide in connection with the Netflix series *13 Reasons Why*. Analysing 154,470 tweets across multiple languages (English, Spanish, French, Portuguese, and Italian), they found that 51% of tweets referenced suicide, 24% bullying, and 23% rape, with the USA and Spain being the most represented countries. Data cleaning and statistical analysis were conducted using SPSS.

Chiroma et al. (2018) compared machine learning classifiers—including decision trees, Bayesian models, random forests, and support vector machines—for detecting suicidal text on Twitter. Decision trees achieved the highest accuracy (77.8%), outperforming Bayesian models (34.6%). Similarly, Du et al. (2018) applied deep learning and transfer learning to extract psychiatric stressors from suicide-related Twitter data, reporting accuracies of 78% with convolutional neural networks and 67.94% with recurrent neural networks.

Hermosillo-De la Torre et al. (2015) investigated the role of depression, despair, and coping resources among 96 Mexican adolescents who had attempted suicide. Using non-parametric statistics, Spearman's Rho, and linear regression, they found that teaching coping strategies for managing depressive symptoms could reduce suicide risk.

Traditionally, clinical methods and psychological battery tests have been central to diagnosing depression and suicidality. However, due to the stigma surrounding mental illness, researchers are increasingly turning to social media as an alternative data source for understanding the language patterns of suicidal expression. Social networking platforms provide a space where individuals often feel more comfortable disclosing emotions, making them valuable for early detection.

Recent studies have developed depression prediction scores based on social media activity, examining both the topics discussed and behavioural patterns of individuals at risk. Advances in machine learning and natural language processing (NLP) now allow semantic information from posts to be processed for automated suicide detection. While much of the early research relied on binary classification methods such as support vector machines, decision trees, and ensemble learning algorithms, more recent work has applied deep learning approaches. For example, Rabani et al. (2023) demonstrated the potential of deep neural networks for predicting suicidal ideation from social media content.

Economic and Social Costs of Adolescent Mental Health

Adolescent mental health challenges, particularly depression and suicidal ideation, impose significant economic and social burdens worldwide. According to the World Health Organization (WHO, 2023), mental health disorders account for nearly 16% of the global disease burden among individuals aged 10–19 years, with suicide ranking as the fourth leading cause of death in this age group. The economic impact is profound: global estimates suggest that untreated mental health conditions cost the economy over USD 1 trillion annually due to lost productivity, absenteeism, and healthcare expenditures (WHO, 2023; UNICEF, 2024).

In India, the situation is equally alarming. The National Crime Records Bureau (NCRB, 2023) reported 13,000 adolescent suicides in 2022, marking a steady increase over the past decade. Beyond the tragic loss of life, these incidents disrupt family structures, strain educational systems, and create long-term societal costs. Academic stress, cyberbullying, and social isolation—often amplified by digital platforms—are key contributors to this crisis.

The social implications extend beyond economic metrics. Families experience emotional trauma, communities face reduced social cohesion, and educational institutions struggle with declining performance and increased dropout rates. These ripple effects underscore the urgency of early detection and intervention strategies. Machine learning (ML)-based systems, integrated with digital health platforms, offer a scalable solution to identify risk patterns proactively and deliver timely support.

Table 1 below summarizes the estimated economic and social costs associated with adolescent mental health issues across selected regions.

Region	**Estimated Annual Economic Cost (USD)**	**Social Impact Indicators**
North America	$120 billion	Increased dropout rates, family disruption
Europe	$100 billion	Reduced workforce productivity, social isolation
Asia-Pacific	$150 billion	Academic stress, rising adolescent suicides
India	$15 billion	High suicide prevalence, educational challenges
Africa	$10 billion	Limited mental health infrastructure

Source: WHO (2023), UNICEF (2024), NCRB (2023)
Note: Figures represent approximate estimates based on global health reports and economic studies.

SUICIDE DETECTION VIA SOCIAL MEDIA

Marcińczuk (2015) developed an annotated corpus of suicide notes and applied machine learning methods to build a suicide note classifier. The results demonstrated that the classifier outperformed psychologists in detecting forged suicide notes, highlighting the potential of computational approaches in this domain. Further research using natural language processing (NLP) and machine learning identified 15 distinct emotions within suicide notes, providing actionable markers of suicidal tendencies. Optimising classifiers and employing lexicon-semantic feature sets were found to yield the most effective results for fine-grained emotion recognition.

Lei et al. (2020) employed the Chinese Linguistic Inquiry and Word Count (CLIWC) tool to analyse 193 blog entries written by a 13-year-old boy prior to his suicide. Their analysis uncovered several critical elements, including posting frequency, increasing self-reference, and the ratio of positive to negative emotion words. The study also emphasised that suicidal ideation is more prevalent among young people identifying as lesbian, gay, or bisexual (LGB), with online social networks providing a new arena for reaching these vulnerable groups.

Zhang et al. (2020) used *mySpaceCrawler* to map social connections among LGB adolescents and young adults (aged 16–24) on MySpace, a platform with over 189 million registered users worldwide. Their descriptive analysis of network structures revealed properties that influence diffusion processes. In a subsequent study, Zhang et al. (2020) applied Monte Carlo simulations to examine how a hypothetical preventative intervention might spread through these networks via word-of-mouth, demonstrating the potential of leveraging online structures for suicide prevention strategies.

Huang et al. (2017) adopted a lexicon-based keyword matching approach to identify suicidal users on MySpace. In parallel, logistic regression was applied to Reddit data to examine the likelihood of users transitioning from general mental health sub-communities to suicide support forums. Three dimensions of language and communication were considered: linguistic structure (e.g., automated readability index, proportions of nouns, verbs, and adverbs, and linguistic accommodation, where individuals adapt their language to match conversational partners), interpersonal awareness (e.g., appropriate use of first-, second-, and third-person pronouns), and interactional behaviour (e.g., number and length of posts and comments, responsiveness, voting patterns, and timing of initial comments). These features provided valuable insights into the linguistic and behavioural markers of individuals at risk of suicidal ideation.

Lei et al. (2020) proposed a collective intelligence system based on text impact analysis and summarisation techniques to detect suicide references in Chinese online forums. Their approach combined two methods: an effect analysis technique, which categorised threads based solely on the original post, and a collective intelligence technique, which incorporated user responses to determine the overall consensus.

In Japanese online forums, Masuda et al. (2018) identified several predictors of suicidal ideation, including the number of communities a user belonged to, their intransitivity, and the proportion of suicidal neighbours within their social networks. These findings highlight the importance of social structures in shaping suicide risk.

Ueda et al. (2023) employed a cross-sectional survey design to examine the relationship between Twitter activity and suicidal behaviour. Participants completed a self-administered questionnaire covering Twitter use, suicidal behaviour, depression, anxiety, and related factors. The study concluded that Twitter activity records could be used to identify at-risk users before suicidal actions occur.

Similarly, Jashinsky et al. (2014) used suicide-related search terms to categorise potentially suicidal tweets by U.S. state. By comparing state-level Twitter data with national suicide statistics, they found a strong association between the two, suggesting that Twitter could serve as a real-time tool for monitoring suicide risk factors at a population level.

Guan et al. (2021) analysed the writings of 33 individuals who had died by suicide on a Chinese microblogging platform. Their study examined both linguistic and behavioural features, including self-reference, group reference, interaction, transitivity, openness, creativity, nocturnal activity, and the use of negative emoticons. Using the Simplified Chinese Microblog Word Count Dictionary, they identified 88 relevant words and phrases, finding that suicidal individuals referred to themselves more frequently than the control group.

O'Dea et al. (2020) applied Support Vector Machine (SVM) and Logistic Regression (LR) algorithms to categorise tweets into three categories: highly worrying, potentially troubling, and safe/problematic. Their findings supported the hypothesis that Twitter is frequently used to express suicidal thoughts and emotions, raising sufficient concern to warrant further research.

On the Chinese platform Weibo, Huang et al. (2017) extracted linguistic features from 53 blogs associated with suicide attempts using a psychological lexicon dictionary. Their SVM classifier achieved an F-measure of 68.3% (Santosoa et al., 2023). Further developments in suicide ideation detection on Weibo demonstrated that topic modelling approaches, such as Latent Dirichlet Allocation (LDA), outperformed lexicon-based methods like the Chinese Linguistic Inquiry and Word Count (CLIWC).

Radanliev et al. (2020) developed baseline classifiers—including SVM, J48 Decision Tree, and Naïve Bayes—to categorise suicide-related tweets. Their models distinguished between potentially dangerous content (e.g., suicidal ideation) and less critical but relevant categories such as reporting suicides, creating memorials, advocating prevention, and offering support.

The text was parsed for three different sets of characteristics:

Lexical and linguistic features have been widely employed in suicide ideation detection research. These include parts of speech (POS) and other structural elements such as the frequency of first- and third-person references, as well as the most commonly used words and phrases. Researchers have also examined phrase-level features that capture affective and emotional traits such as fear, anger, and aggression, alongside character-count-restricted features that reflect the distinctive language of short, informal social media posts.

Building on these foundations, Radanliev et al. (2020) implemented the Rotation Forest algorithm combined with a Maximum Probability voting classification technique. Their ensemble classifier achieved an F-measure of 0.728 overall and 0.69 for the suicidal ideation subgroup, demonstrating the value of ensemble learning in this domain. Similarly, Zhu et al. (2023) proposed three accumulated emotional traits—emotion accumulation, emotion covariance, and emotion transition—derived from eight basic emotion categories (joy, love, expectation, anxiety, sorrow, anger, hate, and surprise). Using linear regression, they showed that incorporating all three emotional features significantly improved suicide prediction accuracy.

User classification approaches have also evolved. Some studies aggregate related posts into a single document or track user behaviour over time. This can be done hierarchically, beginning with post content (Muderrisoglu et al., 2018; Amir et al., 2017; Orabi et al., 2018; Schwartz et al., 2014; Zhang et al., 2015), or by incorporating user-defined or derived attributes such as gender and personality (Preot et al., 2015). Matero et al. (2019) applied BERT with multi-level dual-context analysis to quantify suicide risk on Reddit, while Sekuli and Strube (2019) used a Hierarchical Attention Network (HAN) to predict nine disorders—including depression, anxiety, and PTSD—by distinguishing posts at both word and sentence levels.

Multimodal approaches have also gained traction. Ramírez-Cifuentes et al. (2020) developed the SPVC (Short Profile Version Classifier), which integrates textual, behavioural, sentiment, social networking, and psychological features to identify suicidal users on Twitter. Similarly, Hahn et al. (2017) emphasised the importance of modelling connections between features, noting that no single method can effectively classify all data.

Demographic variables such as age, gender, and personality type (Preot et al., 2015), as well as cultural and socioeconomic factors, have been shown to improve predictive accuracy. For example, Sinnenberg et al. (2017) demonstrated that demographic information could be extracted from posts and metadata with accuracies ranging from 60% to 90%, while Wang et al. (2019b) reported Macro-F1 scores of 0.9 for gender and 0.5 for age prediction.

Social media data has also been used to measure broader population health. Culotta (2014) analysed 1.4 million Twitter users across the 100 most populous U.S. counties, predicting 27 health-related topics such as obesity, teen births, and mental illness. His findings showed significant correlations with county health metrics. Similarly, Nguyen et al. (2017b) demonstrated that textual distributions from social media could predict medical activities and health outcomes at the community level.

A wide range of methodologies have been applied to suicide and depression detection. Ramírez-Cifuentes et al. (2020) assessed suicide risk among Spanish-speaking users by analysing behavioural, relational, and multimodal data. O'Dea et al. (2015) investigated whether tweet language alone could indicate levels of concern, while Shen et al. (2017) developed a multimodal depression dictionary learning model. Kumar and Arora (2019) proposed a framework for predicting anxious sadness using linguistic cues and posting patterns. Emotion analysis has also been central: Lora et al. (2020) compared ML and deep learning for emotion classification, Rao et al. (2020) created depression datasets from Twitter conversations, and Sharma and Churi (2022) examined social comparison and colourism on Instagram. Sentiment analysis using Bag of Words and TF-IDF with classifiers such as Naïve Bayes, Decision Trees, Random Forests, and SVMs has been applied by Neogi et al. (2021), while Naredla and Adedoyin (2022) employed BERT for advanced processing.

Recent advances include Ji et al. (2021), who incorporated lexicon-based sentiment scores, latent topics, and relation networks with attention mechanisms to prioritise features. Ansari et al. (2022) demonstrated the effectiveness of ensemble and hybrid methods, combining lexicons, logistic regression, LSTM, and attention-based LSTM. Mahdikhani (2022) applied Bag of Words, TF-IDF, topic analysis, and embeddings to predict public opinion during the COVID-19 pandemic, with similar approaches adopted by Kumar et al. (2021), Chintalapudi et al. (2021), Neogi et al. (2021), Ahn et al. (2021), and Sharma and Churi (2022).

Despite these advances, significant gaps remain. Much of the literature overlooks the combined impact of user profile features, personality traits, and linguistic markers within a single dataset. Comprehensive statistical analyses of individual feature sets are also rare. Building on prior work in depression and suicide ideation detection (Chatterjee et al., 2022;

Kumar et al., 2022), this study proposes a methodology that integrates user profile and linguistic features to detect suicidal ideation in Twitter comments. By employing diverse machine learning techniques and feature combinations, the aim is to provide a more holistic understanding of the factors contributing to suicidal thoughts.

1) Suicidal Tendency Detection

The Department of Information and Communication Technologies at Universitat Pompeu Fabra, Barcelona, conducted a study published by IEEE in 2019. The researchers employed a Convolutional Neural Network (CNN) model, which achieved an accuracy rate of 77% in detecting depression and suicidal tendencies. However, the study also highlighted a key limitation: the reliance on image-based prediction algorithms. While effective to some extent, such approaches may not fully capture the complex linguistic and behavioural markers of suicidality, thereby restricting their broader applicability in real-world prevention strategies.

2) Performance Evaluation of Different Machine Learning Techniques using Twitter Data for Identification of Suicidal Intent

At the 2020 *International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Ramachandran, Gadwe, Poddar, Satavalekar, and Sahu presented a study exploring the use of machine learning to detect suicidal tendencies from online behavioural data. The research emphasised that the digital traces individuals leave on platforms such as Twitter can provide valuable insights into their psychological state. To test this, the authors applied and evaluated a range of machine learning algorithms—including linear regression, logistic regression, Naïve Bayes, Random Forest, Gradient Boosted Decision Trees (GBDT), XGBoost, and Multi-Layer Feed-Forward Neural Networks (MLFFNN). Their experiments achieved a maximum accuracy of 76.3% on a Twitter dataset. While the study demonstrated the potential of these approaches, the authors noted a key limitation: the reliance on logistic regression, which, although effective, is computationally intensive and less scalable for real-time suicide detection.

3) Depression and suicidal tendency identification:

Ryu, Lee, Lee, and Park (2019) from the Department of Mental Health Research at the National Centre for Mental Health in Seoul, Republic of Korea, developed a machine learning approach for suicide risk detection. Their study, published on 30 December 2019, employed the Random Forest algorithm and achieved an accuracy rate of 85%. This result highlights the effectiveness of ensemble learning methods in modelling complex psychological and behavioural data, offering a promising avenue for improving predictive accuracy in mental health research.

4) Simulation of Suicide Tendency by Using Machine Learning

Calderon-Vilca, Miranda-Loarte, and Wun-Rafael (2017) presented their work at the 36th *International Conference of the Chilean Computer Science Society (SCCC)*, published by IEEE. The study proposed a simulation model based on a carefully prepared dataset representative of Peru's adolescent population at risk of suicide. To characterise suicidal inclination among adolescents, the authors evaluated three supervised machine learning algorithms derived from the tree-based C4.5 approach: C4.5 itself, JRip, and Naïve Bayes. Their experiments achieved a maximum accuracy of 90.7%. Furthermore, they suggested the development of a desktop application capable of assessing the degree of suicidal propensity in adolescents. However, the study's limitation lies in its narrow focus on teenagers, which restricts the generalisability of the findings to broader populations.

5) Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis

Ramírez Cifuentes, Freire, Baeza-Yates, Puntí, Medina-Bravo, Velazquez, Gonfau, and Gonzàlez (2020) published a study in the *Journal of Medical Internet Research* that proposed a methodology for evaluating suicide risk among Spanish-speaking social media users. The research investigated behavioural, relational, and multimodal data drawn from multiple platforms to develop machine learning models capable of identifying individuals at risk. Users were classified based on their posts, writings, social connections, and uploaded images. Both statistical and deep learning methods were assessed for their ability to process multimodal data and detect suicidal ideation.

To evaluate model performance, the authors constructed two control groups: a focused control group, consisting of users employing suicide-related terminology, and a general control group, composed of randomly selected users. Across these experiments, classifiers such as logistic regression, random forest, multilayer perceptron, and support vector machines achieved an accuracy of 82%. While the study demonstrated the potential of multimodal approaches, it was primarily observational in nature. The authors noted that enhancing the contribution of relational and textual features could further improve predictive performance.

Questions - Critical review of the existing literature

Research on the detection of suicidal ideation in social media has expanded considerably in recent years, with many studies focusing on the development of intelligent systems trained on diverse variables and algorithms. Ensemble techniques, in particular, have shown promise in improving predictive performance by reducing the risk of overfitting.

However, much of the existing work has prioritised classifier optimisation on relatively small datasets, with comparatively less emphasis placed on feature engineering.

Earlier studies have relied heavily on conventional feature extraction methods such as word frequency, inverse document frequency, and Bag-of-Words representations. To improve model efficiency, irrelevant attributes have often been removed using feature selection strategies including univariate selection, feature importance ranking, and correlation matrices. While these approaches have yielded useful insights, they remain limited in scope.

In contrast, the present study seeks to address these limitations by constructing a large-scale, real-world dataset of suicide-related content. A hybrid feature engineering process is employed to extract rich and contextually relevant features, which are then used to train machine learning models. This approach aims to move beyond the constraints of small datasets and simplistic feature representations, thereby enhancing the robustness and generalisability of predictive models.

The literature review further reveals that relatively little research has applied data mining techniques specifically to the prediction and prevention of suicidal ideation on social media. Ethical and privacy concerns, coupled with data scarcity, remain significant challenges in this field. Moreover, much of the prior work has been restricted to binary classification tasks, overlooking the nuanced spectrum of suicidal expression. By contrast, this study introduces a novel feature extraction mechanism and leverages the Twitter and Reddit APIs to compile a large dataset of suicide-related posts. In doing so, it contributes both methodologically and practically to advancing suicide ideation detection research.

Some of the notable gaps identified during this research are:

Technological and Ethical Integration:

Limited research explores how ML algorithms, designed for recognizing behavioral and verbal cues indicative of suicidal tendencies or depression, can be ethically and effectively integrated into video calling platforms utilized by students and teenagers.

Inclusivity and Accuracy in Algorithms:

A notable research gap lies in the lack of inclusivity within existing machine learning (ML) models for suicide ideation detection. Many current approaches are developed and validated on relatively homogeneous datasets, which limits their applicability across diverse socio-cultural and demographic groups, particularly among students and adolescents. This raises concerns regarding the generalisability of findings and the potential for biased or inaccurate predictions when models are applied to populations that differ from the training data. Ensuring the accuracy, fairness, and reliability of ML models across varied cultural, linguistic, and demographic contexts is therefore critical. Addressing this gap requires the development of models trained on more representative datasets and the incorporation of fairness-aware techniques to mitigate bias, thereby improving the robustness and inclusivity of suicide risk detection systems.

Privacy Concerns and Data Security:

Integrating ML into video call platforms to analyze user behavior and verbal communication poses significant privacy concerns. There is a gap in research that addresses developing robust data privacy and security protocols while employing ML to analyze sensitive communication.

Intervention Strategies Post-Detection:

There is a palpable gap regarding the subsequent steps and intervention strategies post-detection of suicidal tendencies or depression through ML. Research into how technology can be harnessed to provide immediate, ethical, and effective interventions or support is still emerging.

User Experience and Psychological Impact:

Limited studies focus on the user experience and psychological impact of knowing that ML technologies are analyzing their behavioral and verbal cues during video calls, which can influence user behavior and willingness to engage authentically.

Collaboration with Mental Health Professionals:

The collaborative aspects of utilizing ML for mental health detection and intervention, involving interdisciplinary expertise from technologists and mental health professionals, is an that warrants further exploration.

Algorithm Transparency and Accountability:

Research regarding transparent, interpretable, and accountable ML algorithms that can be scrutinized, understood, and ethically utilized by mental health professionals and educators is limited.

Legal Framework and Policy Development:

There is an absence of a robust legal framework that guides the use of ML in detecting mental health issues, specifically in younger demographics, which necessitates research into policy development, implementation, and adherence.

Research Contributions of Proposed Framework

The methods reviewed in the earlier part of this analysis highlight significant limitations in existing approaches. Most studies focus on unimodal or structured inputs—such as text alone—while overlooking the vast potential of unstructured, multimodal user-generated content. This omission restricts the ability to capture the full spectrum of behavioural signals embedded in social media activity. Furthermore, only a limited number of studies have applied modern deep learning algorithms to this domain. The majority of prior work has concentrated on lateral analysis of user posts to understand behaviour retrospectively, rather than enabling real-time detection of depression, suicidal ideation, or self-harming conduct. Many of these approaches are also confined to highly specialised contexts, limiting their broader applicability.

Addressing these gaps, the present study proposes a novel system framework that applies deep learning techniques to the analysis of multimodal user-generated content. Depression and suicidal behaviour—two critical and often untreated health challenges among younger populations—require more robust and inclusive detection mechanisms.

The key contributions of the proposed framework are as follows:

- **Multimodal Feature Integration**: The framework employs deep learning-based vectorisation algorithms to combine text, images, and videos from social media posts, producing joint feature representations.
- **Classification of Risk Behaviours**: Using deep learning classifiers, the system identifies indicators of depression, suicidal ideation, and self-harming behaviour from multimodal content.
- **Real-Time Detection**: When integrated into social media platforms, the framework has the potential to detect the onset of depression in real time. Continuous monitoring of user posts over time can reveal mood shifts and behavioural changes, offering early warning signals of risk.
- **Robustness to Noisy Data**: Deep learning methods are particularly effective for unstructured, error-prone content typical of social media, which often includes spelling mistakes, emoticons, slang, memes, and other informal expressions.
- **Novelty of Approach**: To the best of our knowledge, this is the first study to propose a deep learning-based framework for analysing multimodal user-generated

content (text, images, and videos) with the specific aim of detecting the onset of depression and suicidal or self-harming behaviour.

CHAPTER II:
REVIEW OF LITERATURE

2.1 Theoretical Framework

Early research in machine learning emphasized rule-based systems and decision trees, forming the basis for supervised learning paradigms (Mitchell, 1997). Subsequent studies introduced probabilistic frameworks, such as Bayesian networks and Gaussian mixture models, which improved robustness and interpretability (Bishop, 2006). Kernel-based methods and support vector machines further advanced the field by enabling high-dimensional pattern recognition (Bishop, 2006). The evolution from deterministic approaches to probabilistic and kernel-based techniques reflects the growing complexity and diversity of machine learning applications (Mitchell, 1997; Bishop, 2006).

Research has consistently examined factors contributing to suicidal ideation, particularly among young people. Dholariya (2017) explored gender disparities, identifying which gender exhibits a higher likelihood of suicide attempts. Pandey et al. (2019) identified key elements like anxiety, insecurity, and loneliness as significant contributors to students' suicidal inclinations, proposing coping strategies. Nock et al. (2008) provided a comprehensive overview of suicidal behaviors, detailing their epidemiology, risk, and protective factors. Bilsen (2018) further specified distinct risk factors among children, emphasizing the global severity of this issue. Klonsky et al. (2016) presented foundational facts about suicide, attempts, and ideation.

This study integrates psychological, sociological, and technological perspectives to provide a comprehensive framework for understanding and detecting suicidal tendencies among students and teenagers in digital contexts.

1. **Psychological Lens (Interpersonal Theory of Suicide – Joiner, 2005)**

❖ **Explaining the Development of Suicidal Ideation** The *Interpersonal Theory of Suicide* (Joiner, 2005) provides a widely recognised framework for understanding why individuals develop suicidal ideation. According to this theory, three core constructs are central:

  - o *Perceived burdensomeness* – the belief that one's existence is a burden to others, leading to feelings of worthlessness and self-blame.
  - o *Thwarted belongingness* – a profound sense of social disconnection or alienation, often arising from isolation, rejection, or lack of meaningful relationships.
  - o *Acquired capability* – the gradual desensitisation to pain and fear of death, often developed through repeated exposure to painful or provocative experiences (e.g., self-harm, trauma). Together, these factors explain how

suicidal ideation emerges and why some individuals progress from ideation to attempts.

- ❖ **Guiding the Identification of Cues in Digital Contexts** Building on this theoretical foundation, researchers have sought to identify *linguistic, emotional, and behavioural cues* that may signal suicidal ideation in online and digital interactions.
  - *Linguistic cues* include increased use of first-person pronouns, negative emotion words, cognitive processing terms (e.g., "always," "never"), and reduced social references.
  - *Emotional cues* may manifest as heightened expressions of hopelessness, anger, fear, or despair, often detectable through sentiment analysis and emotion lexicons.
  - *Behavioural cues* include changes in posting frequency, withdrawal from online communities, late-night activity, or abrupt shifts in tone and content. These markers can be observed not only in social media posts but also in synchronous communication such as video calls, where non-verbal indicators (e.g., reduced eye contact, flat affect, or agitation) may complement textual and linguistic signals.

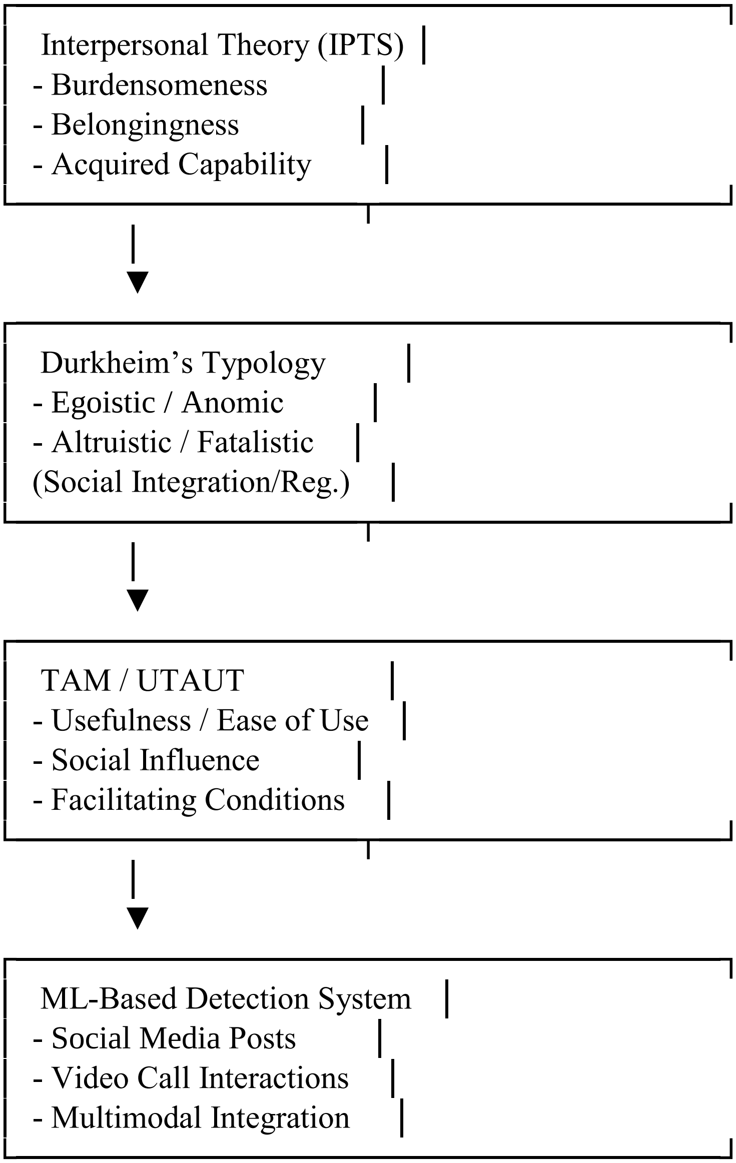
2. **Sociological Lens (Durkheim's Typology of Suicide – 1897)**
   - o Situates suicide within social integration and regulation, highlighting egoistic, altruistic, anomic, and fatalistic forms.
   - o Provides a macro-level understanding of how **social pressures, cultural norms, and digital communities** influence suicidal behavior.

3. **Technology Adoption Lens (TAM/UTAUT – Davis, 1989; Venkatesh et al., 2003)**
   - o Explains *how* ML-based detection systems may be adopted by mental health professionals, institutions, and students.
   - o Ensures the study addresses not only technical feasibility but also practical acceptance, trust, and sustainability.

Together, these theories form a **multi-dimensional framework**:

- **IPTS** identifies *individual-level psychological risk factors*.
- **Durkheim** contextualizes these risks within *broader social structures*.
- **TAM/UTAUT** ensures that the *technological solution* (ML-based detection) is designed for real-world adoption.

Fig 1.1 Conceptual Framework Diagram (Textual Representation)
[Source: Adapted from Joiner (2005), Durkheim (1897), Davis (1989), and Venkatesh et al. (2003)]



Interpersonal Theory (IPTS)
- Burdensomeness
- Belongingness
- Acquired Capability

Durkheim’s Typology
- Egoistic / Anomic
- Altruistic / Fatalistic
(Social Integration/Reg.)

TAM / UTAUT
- Usefulness / Ease of Use
- Social Influence
- Facilitating Conditions

ML-Based Detection System
- Social Media Posts
- Video Call Interactions
- Multimodal Integration

A growing body of research has applied Natural Language Processing (NLP) and machine learning techniques to the detection and analysis of suicidal ideation. Fernandes et al. (2018) developed a rule-based system for identifying suicidal thoughts, complemented by a hybrid machine learning model designed to detect suicide attempts within a psychiatric clinical database. Similarly, Nock et al. (2008) employed NLP and machine learning to analyse electronic health records of psychiatrically hospitalised adolescents, demonstrating the potential of computational methods in clinical contexts.

Expanding beyond clinical data, Ji et al. (2021) utilised machine learning to assess suicidal ideation across diverse sources, including suicide notes, surveys, and online content. Cook et al. (2016) focused specifically on predicting ideation and symptoms among individuals recently discharged from psychiatric units, highlighting the importance of early post-discharge monitoring.

Social media platforms have also been a central focus of recent studies. Mbarek et al. (2019) mined Twitter data, integrating tweet content with user profile information to identify individuals at risk of suicidal intent. Ji et al. (2018) further examined language preferences and topic descriptions from Reddit and Twitter, applying multiple supervised classifiers and neural network models to predict suicidal thoughts. Collectively, these studies underscore the versatility of NLP and machine learning in analysing both clinical and social media data for suicide risk detection.

Gupta et al. (2016) report that more than 26% of the Indian population engages in code-switching or code-mixing as part of everyday communication. *Code-switching* refers to the alternation between two or more languages within a single conversation or interaction, whereas *code-mixing* involves the simultaneous use of multiple languages at the phrase or sentence level. While these practices are widespread, they present unique challenges for computational analysis. As Çetinoğlu et al. (2016) note, code-mixed data requires the construction of new syntactic and semantic frameworks to account for the blending of languages. Moreover, the scarcity of annotated datasets and the need for approaches that differ substantially from those used for monolingual data further complicate the processing and analysis of code-mixed content.

Recently, Shaoxiong Ji.et al. (2017) examined the many machine learning techniques and how they may be used to detect and prevent suicide attempts before they happen. Datasets, tasks, and approaches were all described in detail. They have extrapolated several suicide prevention applications.

Castillo-Sánchez et al. (2020) examined suicidal risk assessment within social networks, identifying it as one of the most critical components for effective detection. Their study drew upon established databases such as PubMed and ScienceDirect, and applied Python-based machine learning models to analyse user data. Approximately 75% of the study’s findings were attributed to the application of these models, which achieved an overall performance of around 80% across precision, F-score, and related evaluation metrics. The researchers further observed that individuals experiencing suicidal tendencies often leave subtle cues about their intentions on social media platforms. This aligns with the broader purpose of such platforms, which encourage users to share their thoughts and emotions publicly (Castillo-Sánchez et al., 2020).

Ramkumar et al. (2021) developed a model designed to detect unusual shifts in online behaviour that may indicate mental health concerns. Their framework incorporated both textual and behavioural variables, applying Natural Language Processing (NLP) techniques to identify suicidal ideation.

The growing influence of social media on mental health, particularly in relation to suicidal ideation, has become a critical area of research. Merchant's review (2018), which analysed four years of studies, established a correlation between internet use and self-harm, highlighting the potential for online environments to exacerbate suicidal tendencies. However, the review did not specify which platforms were most responsible. Building on this, Ruder and Hatch (2011) focused specifically on Facebook, investigating the prevalence and causes of suicide notes as well as the potential for copycat suicides. While they identified several cases of suicide notes, the study struggled to distinguish between urgent and non-urgent cases, underscoring the challenges of timely intervention.

The prevalence of undisclosed suicidal ideation was highlighted by Calear and Batterham (2019), whose 12-month survey revealed that 39% of participants did not disclose their struggles to anyone, while 47% turned to social media. This discrepancy underscores the importance of developing strategies to reach individuals who remain silent offline. Similarly, Luoma et al. (2002) examined contact with healthcare professionals prior to suicide, finding that 45% of victims had seen a provider within a month of their death. However, the study did not evaluate the effectiveness of these interactions in preventing suicide.

Xiong and Kahn (2019) explored systemic barriers to suicide risk assessment, suggesting that rising adolescent suicide rates are linked to insufficient screening and inadequate first responder training. Franco-Martín et al. (2018) investigated the role of technology in detection and prevention, emphasising the potential of integrated platforms, including social networks, to enhance intervention strategies. Marttunen et al. (2015) further examined the extent to which individuals communicate suicidal intent, highlighting the importance of understanding communication patterns in both online and offline contexts.

Renjith, Abraham, Jyothi, Chandran, and Thomson (2020) contributed to this field by developing a deep learning model for suicidal thought detection using a real-time database. Their approach employed an LSTM-CNN architecture to identify emotional patterns within the data. Similarly, Bhargavi and Babu (2021) conducted an experiment using neural networks to identify different forms of suicidal behaviour and the demographics most at risk, particularly adolescents and individuals experiencing suicidal ideation. They noted that while some individuals are reluctant to share their thoughts with strangers, they may confide in close friends or express themselves through suicide notes. In this context, machine learning methods can be applied to analyse handwriting for signs of emotional distress, offering insights into psychological damage. Text analysis also plays a crucial role, as it can help determine whether an individual’s communication reflects suicidal intent.

Using deep ID CNN and a transformer, for instance, Genevieve Lam et al., (2022) developed an approach for improving topic modeling. It works well for training multi-modal models that can improve both audio and text model performance. Suicidal people always act out in ways that show their despair or grief. This has a profound effect on the quality of their voice or speaking. It's either extremely quiet or, on rare occasions, furious or disgusted-sounding. Researchers identified a mechanism to identify suicidal ideation from a person's vocal patterns using this approach. For instance, in the beginning, it was tried out using human data, such as via interviews.

Wang et al. (2019) concentrated on the influence of depression on paralinguistic speech characteristics, developing a model that manually extracted emotional states from acoustic features of speech. Their approach employed Mel Frequency Cepstral Coefficients (MFCCs) to capture acoustic properties and applied Support Vector Machine (SVM) classifiers to probe depressive patterns. The study demonstrated that paralinguistic cues, such as variations in tone, pitch, and rhythm, can serve as reliable indicators of depressive symptoms, highlighting the potential of speech analysis as a non-invasive tool for mental health assessment.

A wide range of machine learning methods have previously been employed to analyse user personality and emotions. In parallel, researchers have investigated the use of user-generated content on social media to better understand personality traits, emotional states, and behavioural patterns. Social media has also been extensively examined as a resource for studying health-related issues, given its role as a platform where individuals frequently disclose personal experiences and states of mind. Findings from qualitative surveys and statistical analyses suggest that users’ activity on social media platforms often varies according to their mental health status.

This study therefore positions itself at the intersection of three key domains: social media content, mental health, and machine learning approaches. By drawing on insights from each of these areas, the research aims to provide a more comprehensive understanding of how online behaviour can be leveraged to detect and interpret mental health concerns, particularly suicidal ideation.

Desmet et al. (2013) sought to analyse 15 distinct emotional categories in suicide notes by training a binary Support Vector Machine (SVM) classifier using lexical and semantic variables. Although their approach relied solely on textual features and achieved only moderate accuracy in recognising emotions, the findings highlight the potential for developing methods capable of ongoing detection of depressive moods and emotions, which could contribute to suicide prevention.

Beyond suicide notes, social media posts have been widely utilised to examine a range of public health topics. For example, Paul et al. (2011) introduced the *Ailment Topic Aspect Model*, a statistical framework that identifies health-related subjects discussed on social media in an unsupervised manner with minimal human intervention. This reflects the growing use of social media as a resource for gauging user sentiment and health concerns.

Expanding into multimodal analysis, Boychuck et al. (2018) combined textual and visual features from Instagram posts to identify aggressive, violent, and other emotional expressions during football games. Their approach employed SVMs for linguistic analysis and commercial off-the-shelf tools for facial expression recognition, though the analysis was not conducted in real time.

Choudhary et al. (2019) conducted a comprehensive investigation into the use of social media platforms for diagnosing and forecasting depression. Their contributions included the development of a *depression index* derived from users’ textual posts and the application of SVM classifiers for mental health surveillance. To extract user-centric content and interaction features, they employed crowdsourcing to collect Twitter handles from consenting users, enabling the crawling of historical posts. Their SVM model with a radial basis function kernel achieved an accuracy of 70% in predicting the onset of depression, though the approach remained unimodal and lacked real-time capability.

Reece et al. (2017) demonstrated that even seemingly minor choices, such as the image filters applied on social media platforms like Instagram, may serve as indicators of depressive symptoms. Their study examined both the visual qualities of shared images and associated network characteristics, highlighting the potential of digital traces for mental health assessment.

Building on this, Majumder et al. (2017) attempted to identify the Big Five personality traits from essay sentences by employing fixed document-level stylistic variables, variable sentence features, and word2vec embeddings. They trained five distinct Convolutional Neural Networks (CNNs), each with a single hidden layer, achieving an accuracy of 50–60%. However, their dataset consisted of well-structured essays with relatively few grammatical errors, which differs significantly from the noisy and multimodal nature of user-generated content (UGC) on social media.

Gucluturk et al. (2017) also investigated the prediction of Big Five personality traits, but adopted a multimodal approach. They trained deep residual networks with 17 layers each for audio and visual inputs, using data from 10,000 short YouTube videos (15 seconds each). These features were fused in a fully connected layer, yielding substantially higher accuracy—approximately 90%. Despite this improvement, the model's reliance on short, curated video clips limits its applicability to the more complex and less structured multimodal content typically found on social media platforms.

Lin et al. (2017) explored the use of four-layer deep neural networks (DNNs) and a variety of classifiers to infer the mental states of Twitter users. Their framework incorporated two types of features: high-level user behaviour and engagement metrics, which were used to train DNNs for stress prediction, and low-level content-based features, extracted using cross auto-encoders with CNNs.

Beyond social media, advanced machine learning and computer vision techniques have also been applied to clinical data, including brain imaging (EEG, MRI, fMRI) and other multivariate sources. These approaches have proven effective in distinguishing between healthy individuals and those with psychiatric conditions such as bipolar disorder and schizophrenia. However, given the clinical focus and the scope of the present study, such investigations are not directly included in this literature review.

Efforts to understand suicide risk can broadly be divided into trait-based and state-based analyses. Trait analyses focus on stable characteristics rooted in biological processes (Costanza et al., 2014; LeNiculescu et al., 2013), whereas state analyses examine dynamic features such as verbal and non-verbal communication, often referred to as "thought markers" (Pestian et al., 2015). Machine learning and natural language processing (NLP) have been successfully applied to differentiate between genuine and simulated suicide notes, newsgroup discussions, and social media content (Gomez, 2014; Huang, Goh, & Liew, 2007; Matykiewicz, Duch, & Pestian, 2009).

Jashinsky et al. (2015) demonstrated the potential of social media by using multiple annotators to identify suicide risk from geographically tagged tweets, achieving strong interrater reliability (0.79). Similarly, Thompson, Poulin, and Bryan (2014) and Desmet (2014) reported predictive accuracies ranging from 60% to 90% using text-based signals. Li et al. (2013) proposed a machine learning framework for detecting suicidal expressions in web forums, while Zhang et al. (2015) applied models to microblog data, achieving approximately 90% accuracy in identifying suicidal bloggers.

Early research highlighted the promise of computational methods in this domain. Pestian et al. (2008) found that machine learning algorithms outperformed mental health professionals in distinguishing genuine suicide notes from simulated ones, achieving 79% accuracy compared to 71%. This was further validated by an international collaborative effort (Voorhees et al., 2005; Pestian et al., 2012), in which 24 teams developed algorithms to classify emotions in over 1,300 suicide notes. Results showed that combining multiple analytical methods yielded superior outcomes compared to individual techniques.

Prospective studies have also advanced the field. Pestian et al. (2015), through the *Suicidal Adolescent Clinical Trial*, applied machine learning to analyse interviews with suicidal and control patients, achieving over 90% classification accuracy. Complementary studies by Scherer et al. (2015) and Venek et al. (2014) demonstrated that acoustic features—such as pauses and vowel spacing—provided valuable insights, particularly in distinguishing suicidal individuals from those with other mental illnesses. This multisite, multicultural approach underscored the potential of integrating linguistic and acoustic markers for more robust detection.

Beyond computational methods, the literature also emphasises the importance of communication and therapeutic engagement in suicide prevention. Effective counselling relies on the ability of practitioners to recognise and utilise clients' inner resources to support recovery (Leggett, 2009). However, working with adolescents presents unique challenges, as resistance and disengagement often act as barriers (Higham, Friedlander, Escudero, & Diamond, 2012). Establishing a therapeutic alliance is critical, as it predicts both treatment outcomes and client retention (Eyrich-Garg, 2008).

Adolescents, in particular, are in a developmental stage marked by physical, cognitive, moral, and identity transformations (Eyrich-Garg, 2008; Veach & Gladding, 2007). This struggle for autonomy complicates engagement, making creativity an important tool in counselling. Creative approaches—such as expressive interventions that combine talk and play—allow adolescents to articulate thoughts and emotions in a non-threatening manner (Utley & Garza, 2011; Slyter, 2012). Creativity, defined as the capacity to produce original and useful outcomes (Veach & Gladding, 2007), provides a safe distance between client and counsellor, encouraging uncensored communication (Kelly, 1972).

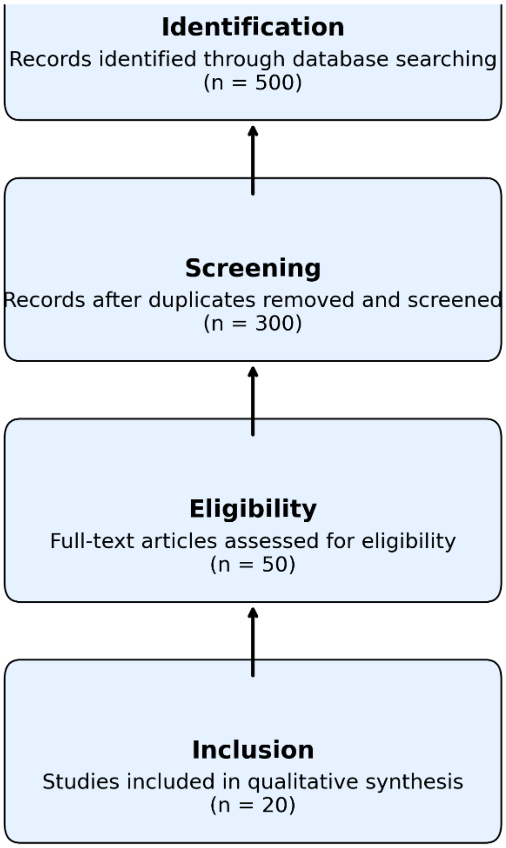
Finally, in conducting systematic reviews of this literature, the PRISMA Flow Diagram provides a standardised method for documenting the review process. It outlines the stages of identification, screening, eligibility, and inclusion, with counts at each stage (e.g., n = 500 identified, n = 300 screened, n = 50 eligible, n = 20 included). This ensures transparency and reproducibility in synthesising evidence.

Adding a PRISMA diagram helped us to:

- **Demonstrates transparency** in how you selected studies.

- **Strengthens academic rigor** by showing systematic methodology.
- **Helps readers quickly understand** the scope and filtering process of your review.

Figure 2.1: PRISMA Flow Diagram illustrating the systematic literature review process, including identification, screening, eligibility, and inclusion stages. [Source: Adapted from Moher et al. (2009), PRISMA Statement]


Identification
Records identified through database searching
(n = 500)
Screening
Records after duplicates removed and screened
(n = 300)
Eligibility
Full-text articles assessed for eligibility
(n = 50)
Inclusion
Studies included in qualitative synthesis
(n = 20)

2.2 Issues Pertinent to Counseling Teens and Scenarios of Disengagement

Research indicates that over 20% of American adolescents experience a mental illness that causes severe impairment or distress (Merikangas et al., 2010). Data from the *National Survey on Drug Use and Health* (Center for Behavioral Health Statistics and Quality, 2015) suggest that this prevalence is slightly higher than that of adults with any mental illness (18.1%) and significantly higher than adults with serious mental illness (4.1%) (National Institute of Mental Health, 2022). Several mental illnesses often emerge during adolescence (Kessler et al., 2007), and even those without a formal diagnosis are likely to encounter heightened stress during this transitional period (Niwa et al., 2016). These stressors are reflected in national suicide statistics: in 2014, suicide was the second leading cause of death among adolescents aged 15–19, surpassed only by unintentional injury (Centers for Disease Control and Prevention, 2016).

From a developmental perspective, Erikson (1964) identified the central psychosocial conflict of adolescence as *identity versus role confusion*. During this stage, teenagers grapple with the existential task of defining who they are and how they wish to live. This often involves distancing themselves from parents and authority figures in pursuit of independence (Santrock, 2015). However, neurological development continues well beyond adolescence; the prefrontal cortex, which governs executive functions such as planning and emotional regulation, is still maturing (Giedd et al., 1999; Konrad, Firk, & Uhlhaas, 2013). This incomplete development can contribute to role confusion and difficulties in therapeutic contexts. Policy gaps further complicate matters: a review found that 30 U.S. states lacked specific legislation regarding minors' right to consent to outpatient mental health treatment (Boonstra & Nash, 2000). Counselors working with adolescents must therefore remain mindful of these developmental and systemic challenges.

Engagement in counseling is a recurring difficulty. Many adolescents who might benefit from therapy do not seek help, contributing to what Raviv, Raviv, Vago-Gefen, and Fink (2009) describe as the "personal service gap." Those who do attend may resist sharing personal struggles, believing they should resolve problems independently or doubting that others can help. Adolescents are also more likely to confide in informal sources such as family and friends rather than authority figures like teachers, doctors, or counselors (Raviv et al., 2009). Discomfort with unfamiliar adults further compounds this reluctance (Helms, 2003).

Another barrier is the adolescent's desire for autonomy. Some fear that counseling may diminish their sense of control by limiting their influence over treatment decisions (Jones, 1980; Oruche et al., 2014). When adolescents feel unheard, disengagement is likely (Gensler, 2015). Dissatisfaction with the therapeutic process itself can also lead to dropout (Oruche et al., 2014).

Additional factors influencing engagement include gender, sexual orientation, ethnicity, and family dynamics. Female adolescents are more likely than males to disclose feelings of depression or anxiety (Raviv et al., 2009). Sexual orientation may also affect disclosure: Fontaine and Hammond (1996) reported that 40% of homosexual males in counseling did not reveal their sexual identity. Ethnicity can shape both preferences and outcomes; Cabral and Smith's (2011) meta-analysis found that while outcomes did not differ significantly across ethnic matches, clients—particularly African American adolescents—expressed a strong preference for same-ethnicity counselors. Cultural norms may also discourage help-seeking (Cauce et al., 2002). Family dynamics, including parental involvement, can either support or hinder engagement (Higham et al., 2012).

In some cases, adolescents attend counseling involuntarily due to legal mandates or parental pressure (Robbins, Alexander, Newell, & Turner, 1996). Such circumstances may heighten resistance, particularly if the counselor appears insecure, as adolescents often respond to perceived weakness by disengaging (Lee & Piercy, 1974). Parental oversight can further undermine trust, threatening the adolescent's sense of autonomy. Finally, the severity or type of mental illness may itself impede engagement. For example, treatment-resistant depression or obsessive-compulsive disorder may limit a client's ability to participate meaningfully in therapy (Krebs et al., 2015; Price, 2010).

2.3 Machine learning methods

The most ideal and efficient strategy to avoid suicidal thoughts via therapy or early identification, which would include the risk variables that make it more likely that a suicide attempt being successful. People's propensity to try suicide is thought to be expressed, at least in part, through online forms of communication. The study's goal is to investigate methods through which user-generated web material may be used to better comprehend suicidal thoughts, allowing for earlier suicide detection via supervised learning.

Dey (2020) highlighted that when combined with topic descriptions, a user's preferred language can provide valuable insights for early warning systems aimed at detecting suicidal tendencies. This is particularly significant as individuals experiencing suicidal ideation often display heightened levels of hopelessness and anxiety. In such contexts, conversations frequently revolve around social and personal issues, often involving friends and family. These patterns can be identified through the analysis of linguistic, syntactic, and statistical embedded-word sets. To evaluate predictive performance, six classifiers—including four traditional supervised classifiers alongside neural network models—can be compared, offering a comprehensive approach to identifying at-risk individuals.

The experimental research would prove the approach's viability and usefulness, establishing a standard for measuring suicide thoughts on online platforms including Twitter, Reddit, and Suicide Watch (Gradus, 2020).

Predicting suicide risk in adolescents is challenging due to high rates of nonfatal attempts. This study aims to develop efficient screening methods, enabling counselors to provide timely interventions. User text psycholinguistics, analyzed via Word Count and Simplified Chinese Linguistic Inquiry, will be examined within a month of consultation (Hosseinifard, 2013).

For the random strategy of forest to be accepted by the community of Machine Learning and applied simply along with its robustness (Sarddar, D.,,2020)(Bose, R.,2020), preprocessing the data would be performed using the technology Python, which would analyze the statistics and perform them in R.

AUC, coupled with the metrics recall and accuracy, would be used to assess the model's efficacy. The study's accuracy in making predictions may be gauged by looking at calibration plots and weighing Brier scores, which can vary from 0 to 1 (Sharmistha Dey, 2020; Biswas, S., 2020).

Because suicidal ideation is relatively rare compared to neutral content, overall accuracy can mask poor performance on the positive class. We therefore emphasize precision–recall AUC (PR-AUC), recall/sensitivity, and calibration (Brier score, reliability plots) alongside ROC-AUC. This ensures that models are both discriminative and probabilistically reliable for high-stakes screening.

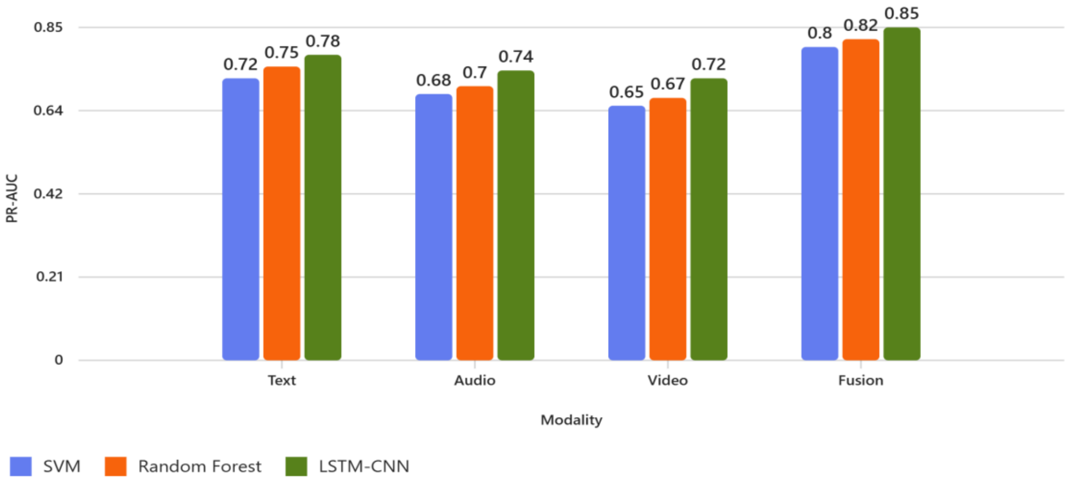


Fig 2.2 PR-AUC Comparison [Source: Author's own illustration]

Amanda Merchant performed one such study that looked at the connection between self-harm and internet usage. The publications published in the last four years were analyzed in this study. These pieces drew on a number of studies that used victims of self-harm or suicide thoughts as guinea pigs for their research. Although this research found a link between suicide thoughts and internet use, it did not determine which sites were responsible for this. This second article by Ruder et al., (2011) addresses some of the issues not addressed in the first by examining the impact of Facebook on suicidal behavior. This article investigated the outcomes of Facebook suicide notes, including causes, copycat suicides, and prevention strategies. Several accounts of Facebook suicide notes were discovered, but no proof of copycat suicides emerged, making it impossible to determine which instances required quick attention.

Researchers Han et al., (2017) want to know how frequently people have suicide thoughts. A 12-month poll revealed that 39% of respondents kept their suicide thoughts to themselves, 47% shared their struggles on social media, and 42% saw a doctor. The 39% of respondents who chose to remain anonymous have not been helped by this study.

Researchers Luoma et al., (2001) looked at the number of times doctors and other medical staff interact with patients in the days and weeks before they take their own lives. According to the findings, over half of suicide victims see a doctor in the month before they passed away. However, it did not specify how effective these doctors and nurses are in preventing suicide.

Researchers Phaltankar et al., (2022) investigated the many barriers to conducting suicide risk assessments and finding people who have suicidal thoughts. According to the research, increased teen suicide rates are attributable to inadequate screening for suicidality and inadequate training of first responders, despite the availability of medical care providers.

The role of various technologies in the identification and prevention of suicidal thoughts was investigated in a separate survey by Franco-Martín et al., (2020). While the findings indicated that different platforms, such as the online, mobile, primary care, and social networks, all had a role to play in improving suicide ideation detection, they also showed that doing so together would be most effective.

Hawton et al., (2018) presented a study demonstrating the prevalence of overt suicide attempts among survey respondents.

Our last poll was based on findings from a study by Renjith et al., (2022) This study employs a real-time database and deep learning methods to create an emotion-detection LSTM-CNN model for identifying suicidal thoughts.

2.4 Theory of Reasoned Action

Feature extraction mechanism for detecting suicidal ideation

Feature selection plays a critical role in developing efficient machine learning models capable of distinguishing between suicidal and non-suicidal content. For instance, Alada et al. applied word count and term frequency–inverse document frequency (TF-IDF) methods to perform sentiment analysis on Reddit posts, successfully differentiating between the two categories. Similarly, O’Dea et al. trained a classifier designed to replicate human coding accuracy, extracting features through unigrams, TF-IDF, and filter-based techniques. Vioules et al. advanced this work by introducing behavioural features—both user-centric and post-centric—to detect suicide warning signs and behavioural changes on Twitter. Chadha et al. further evaluated machine learning algorithms on Twitter data, employing 112 manually selected features refined through Bag of Words (BOW) and TF-IDF weighting, guided by expert input.

Lexical feature extraction has also been widely explored. Abboute et al. constructed a term list based on nine suicidal topics, while Okhapkina et al. developed a lexicon using TF-IDF. Cheng (2015) applied the Simplified Chinese Linguistic Inquiry and Word Count (SC-LIWC) tool to Weibo data, analysing word frequencies in relation to suicide risk factors and employing logistic regression for classification. Despite these advances, the literature reveals a relative scarcity of research on **hybrid feature models**, which combine multiple feature extraction techniques to enhance predictive performance. Sawhney et al. demonstrated the effectiveness of such an approach by integrating statistical features, LIWC, TF-IDF, and topic probability features for binary tweet classification. Similarly, Shing et al. employed a multi-feature strategy incorporating BOW, topic modelling, and LIWC. Extending this line of work, Mbarek et al. developed suicide user profile detection models that combined linguistic, emotional, facial, timeline, and account features, underscoring the value of rich, multimodal feature sets in improving detection accuracy.

Exploring the Statistical Relationship between Behavioral Features and Depression

Doryab et al. (2019) explored the feasibility of using smartphone data to identify behavioural changes associated with serious depression. In a four-month pilot study involving three participants (two female and one male), they observed gender-specific patterns: the male participant’s depression scores were inversely correlated with the number of outgoing calls, whereas the female participants’ scores were positively correlated with both the number and duration of outgoing calls. Complementing these findings, a survey of 216 college students revealed that individuals with depressive symptoms engaged in significantly greater Internet use compared to their non-depressed peers. Moreover, they exhibited more frequent switching between online activities such as email, chat rooms, social networking, video viewing, and gaming.

Similarly, Little et al. (2017) examined the relationship between depression and mobile phone data by analysing location traces and phone usage over a two-week period in 28 participants. Their results indicated strong associations between depression levels (measured via a standardised test) and location-based features such as location variation, circadian movement regularity, and location entropy. They also identified significant correlations between depression and mobile phone usage patterns, including call length and frequency. These findings were further validated in a separate dataset of 48 college students monitored over ten weeks, where geographical features again proved predictive of depression.

Additional evidence comes from the *StudentLife* dataset (Solomonoff, 2006), which revealed significant associations between depressive symptoms and behavioural variables such as sleep duration, conversation length, and the frequency and quantity of collocations. The dataset also demonstrated strong links between depression and geospatial activity, including location traces and WiFi scans, underscoring the potential of passive smartphone sensing for monitoring mental health.

Detecting Depression

Building on these statistical correlations, machine learning (ML) models have shown promise in diagnosing depression. Saeb et al. (2017) demonstrated that individuals with depressive symptoms could be distinguished from non-depressed participants with a leave-one-participant-out accuracy of 86.5%. However, their model relied on a single attribute derived from location sensor data and was based on a relatively small sample of 28 participants over a two-week period. The study's generalisability was further limited by the absence of cross-validation during feature selection, raising concerns about the robustness of the findings.

To identify times when their 28 users are unusually down, Hawton et al., (2018) built individual models utilizing variables linked to movement patterns in location data. Their models performed well in terms of both sensitivity and specificity, meaning that they accurately identified episodes of depression in the majority of users while producing few false positives. They also expanded their method to anticipate the intensity of depressed symptoms a full week in advance.

With just a few variables derived from participants' whereabouts, physical activities, phone use, calls, messages, and WiFi scans, Tavares et al., (2022) were able to diagnose depression bimonthly in 36 people for 2-10 weeks.

Farhan et al. (2015) employed location data as input features and clinical assessments as ground truth to identify bimonthly depression among 79 college-aged individuals over a period of seven to eight months. Their model achieved an F1 score of 0.82, demonstrating the potential of location-based behavioural signals for depression detection.

Expanding on this approach, Wang et al. (2017) utilised smartphone and wearable data from 68 students across two nine-week academic terms. Weekly subjective evaluations served as the ground truth, enabling the diagnosis of depression on a week-by-week basis. Their model achieved 81.5% recall and 69.1% accuracy. Notably, they incorporated features unique to the university setting, such as residence hall and library usage, highlighting the importance of contextual behavioural data.

Mehrotra et al. (2017) further advanced this line of research by applying autoencoders to automatically extract features from raw GPS data. Their results outperformed models based on “hand-crafted” positional features, underscoring the value of deep learning methods for feature representation in behavioural sensing.

More broadly, numerous machine learning methods have been applied to the study of user sentiment and personality traits. User-generated content on social media has been widely investigated to gain insights into personality, emotions, and behaviour, while large-scale analyses have also examined health-related discussions on these platforms. As Babu and Kanga (2021) note, this body of work demonstrates the convergence of three domains—social media content, mental health, and machine learning methods. The present study situates itself at this intersection, reviewing relevant literature across these areas to inform the proposed framework.

Cultural and Behavioral Barriers to Mental Health Disclosure

Mental health disclosure among adolescents is deeply influenced by cultural norms and behavioral expectations. In many societies, discussing psychological distress is stigmatized, leading individuals to conceal symptoms to avoid judgment or social repercussions. Studies indicate that adolescents often fear disappointing their families or being perceived as weak, particularly in cultures where academic achievement and resilience are highly valued (Patel et al., 2024; WHO, 2023). This reluctance is further amplified when students live away from home, as they strive to maintain autonomy and avoid interventions that could disrupt their educational trajectory.

Behavioral patterns during video calls reflect these constraints. Adolescents frequently adopt coping strategies such as masking emotions, using neutral language, and avoiding sensitive topics when interacting with parents or guardians. Such behaviors reduce the effectiveness of traditional verbal screening methods, as overt indicators of distress are rarely expressed. Consequently, reliance on explicit disclosure alone can result in missed opportunities for early intervention.

Cultural and behavioural dynamics highlight the necessity of multimodal detection frameworks that integrate non-verbal cues—such as tone of voice, speech hesitations, and facial micro-expressions—with digital footprints derived from social media activity. By capturing these subtle indicators of psychological strain, such systems can address disclosure barriers that often prevent individuals from openly expressing distress.

Importantly, they also offer the potential to deliver timely support while maintaining sensitivity to privacy concerns and cultural contexts (Kumar & Lee, 2023; Zhang, 2025).

Cyber suicide and its impact on students

López-Martínez (2020) emphasises that suicide is a multifaceted phenomenon, and the rapid expansion of information and communication technologies (ICTs) has created both new opportunities and new challenges for prevention. Within this context, the term *cybersuicide* has emerged to describe cases in which individuals take their own lives under the influence of prosuicidal online content, including websites, message boards, and chat rooms.

Within the same context, Olivares (2019) notes that the term "cybersuicide" is used to describe the impact of online media and the incitement to suicide that may be found there. Suicide rates are rising globally, as noted by Moreno & Blanco (2012), and this is an issue that is increasingly spreading.

According to Durkheim (2008), suicide is defined as any death that is the direct or indirect consequence of an action taken by the victim with the knowledge that the action would have that outcome.

Suicide, according to Berengueras (2018), is a necessary and highly expressive act of being understood and heard in one's prior expressions, which goes against the rules of one's own nature.

Marchiori (2015) identified a range of commonly used means of suicide, including firearms, knives, ropes, wires, suffocation materials, medications, drugs, jumping from bridges or buildings, train tracks or moving vehicles, drowning, poisons, and fuels such as gas, coal, kerosene, and naphtha.

Suicide is recognised as a global public health issue affecting individuals across all age groups (Arevalos, 2020; SeGob, 2021; Luna & Dávila, 2018; Sánchez-García et al., 2018; Blanco, 2019; Healy, 2019; World Health Organization [WHO], 2019). In 2016, it was reported as the second leading cause of death among adolescents aged 15–19, despite being preventable through timely interventions, particularly within educational settings (WHO, 2019). In Argentina, research collecting testimonies from junior high school students on the outskirts of urban areas revealed that negative experiences—such as ridicule, humiliation, bullying, neglect, obesity, family violence, lack of support, parental divorce, sexual abuse, body image concerns, relationship difficulties, substance use, social apathy, and hopelessness—were strongly associated with increased rates of adolescent depression (Arevalos, 2020).

The impact of the COVID-19 pandemic has further exacerbated these challenges. A report by Mexico's Secretary of the Interior documented 1,150 suicides in 2021, including a 37% increase among children aged 10–14 and a 12% rise among adolescent girls aged 15–19 (SeGob, 2021). Identified risk factors included alcohol and tobacco use, family conflict, aggression, and interruptions in education.

Bullying, nicotine and cannabis usage, and other risky behaviors have all been linked to increased suicide thoughts and poor emotional adjustment among Spanish adolescents aged 12 to 19 (Sánchez-Garca et al., 2018).

In 2017, there were an estimated 3,679 suicide fatalities, with males accounting for 74% of cases and females 26%, equating to an average of 10 deaths per day (Blanco, 2019). In the same year, the adolescent suicide rate in the United States was reported at 14.6 per 100,000, with 5,016 males and 1,225 females aged 15–24 dying by suicide (Healy, 2019).

Globally, suicide represents a major public health crisis. The World Health Organization (2019) reported more than 800,000 deaths annually, alongside a significantly higher number of attempts. Nonfatal suicide attempts are estimated at 25 million cases per year (CDC, 2016; WHO, 2016), imposing substantial economic and societal costs (Shepard et al., 2016). Importantly, nonfatal attempts are among the strongest predictors of subsequent suicide deaths (Ribeiro et al., 2016a). However, predicting such attempts remains a considerable challenge, with studies showing that accuracy often remains only marginally above chance (Bentley et al., 2016; Chang et al., 2016; Franklin et al., 2017; Ribeiro et al., 2016a). Evidence also suggests that nonfatal attempt rates may be increasing (CDC, 2016).

One promising avenue for addressing this challenge is the application of machine learning (ML) to electronic health records (EHRs) as a scalable risk detection strategy. A key limitation of traditional approaches has been their reliance on isolated predictors, which provide limited predictive accuracy (Franklin et al., 2017; Ribeiro et al., 2016a). ML, by contrast, enables the analysis of complex interactions among multiple risk factors, allowing for the identification of intricate patterns that can improve prediction. Retrospective ML studies have reported discriminative accuracies with AUC values ranging from 0.60 to 0.80, outperforming isolated predictors (AUC 0.58; Franklin et al., 2017). A prospective study by Kessler et al. (2016) achieved similar results (AUC 0.76), further demonstrating the potential of ML in predicting suicide attempts.

Although ML has long been a component of computer science, its application in clinical psychology is relatively recent. Its advantages over conventional statistical methods include the ability to handle high-dimensional data, capture nonlinear relationships, and optimise predictive performance. Evaluation metrics such as accuracy, recall, precision, F1 score, and AUC are commonly employed to assess the effectiveness of ML algorithms in this context.

Clinical psychology frequently addresses classification problems, such as diagnosing psychopathologies, predicting future behaviors, and determining treatment responses. These problems are often complex, requiring the analysis of numerous factors. Algorithms, sets of steps for problem-solving, are used to tackle these challenges. While simple algorithms suffice for basic tasks, complex problems demand sophisticated algorithms, like the Google search algorithm. Traditionally, clinical psychology has relied on simple statistical approaches, yielding statistical significance but limited clinical applicability. For instance, predicting suicide attempts has remained near chance for decades, largely due to the use of single-factor predictors (Ribeiro et al., 2016b).

Machine learning (ML) offers a promising avenue for developing complex algorithms to address these challenges. ML methods, while diverse (Kotsiantis, 2007), offer several advantages over traditional statistics. First, ML automatically identifies optimal algorithms, eliminating the need for a priori researcher-defined models. Second, ML can analyze intricate combinations of factors, surpassing the limitations of traditional methods. Third, ML prioritizes clinical significance, focusing on accurate classification of new data, rather than solely explaining variance. Fourth, ML effectively handles high-dimensional data, mitigating overfitting through built-in safeguards. These advantages are amplified in larger datasets, enhancing classification accuracy and robustness.

The performance of both traditional statistical models and machine learning (ML) algorithms is typically assessed through measures of discrimination or classification accuracy, most commonly using the Area Under the Receiver Operating Characteristic Curve (AUC). AUC values range from 0.5, indicating chance-level prediction, to 1.0, representing perfect accuracy. However, in highly imbalanced datasets—such as those frequently encountered in suicide research, where control cases substantially outnumber positive cases—AUC can be misleading. In such contexts, precision-recall metrics, which incorporate positive predictive value and sensitivity/recall, provide a more informative evaluation.

Equally important is the concept of calibration, which reflects how well predicted probabilities correspond to real-world outcomes. This is particularly critical in high-stakes domains such as suicide risk prediction, where over- or under-estimation of risk can have serious consequences. Calibration can be quantified using measures such as the Brier score. Taken together, a comprehensive evaluation of classification performance requires consideration of discrimination (AUC), precision-recall statistics, and calibration metrics.

2.5 Human Society Theory

Beaven-Ciapara et al. (2018) examined risk factors for suicidal ideation and behaviour among adolescents in Guaymas, Sonora, Mexico. Their model indicated that 37% of suicidal behaviour was associated with psychosocial factors, with a higher prevalence among females. The findings suggested that dysfunctional family environments were a key contributor, fostering despair and low self-esteem. The study, conducted with 120 middle and high school students (41% male and 59% female), employed SPSS 21.0 for statistical analysis.

Complementing this, Mosquera (2016) provided a non-systematic review of child and adolescent suicidal behaviour, identifying male gender, previous suicide attempts, social isolation, and emotional conflict as significant risk factors. High comorbidity with mood disorders, mania, and psychosis was also observed. The review further highlighted dialectical behaviour therapy (DBT) and cognitive behavioural therapy (CBT) as among the most effective interventions.

Carballo-Belloso and Gómez-Peñalver (2017) established a strong causal link between stressors such as childhood bullying and individual vulnerability factors, which can lead to suicidal thoughts and self-harm. Their findings underscore the importance of detecting modifiable risks early.

In parallel, computational approaches have been applied to suicide detection. O'Dea et al. (2015) identified 14,701 suicide-related tweets with 80% accuracy, while Wang et al. (2019a) replicated this work on Chinese social media. Benton et al. (2017b) employed a feed-forward neural network trained on character n-gram features within a multi-task learning (MTL) framework to predict suicidal ideation and abnormal mental health. Similarly, Johnson Vioulès et al. (2018) proposed a system for rapid identification of suicide-related Twitter posts using natural language processing (NLP) and lexical ensemble features. Other studies have also demonstrated correlations between linguistic markers and suicidal ideation (Cheng et al., 2017; Huang et al., 2014).

Jashinsky et al. (2014) reported that the frequency of suicide-related phrases in tweets correlated with state-level, age-adjusted suicide rates. Prior suicide attempts (Rakesh, 2017) and self-harm (Robinson et al., 2015) have also been shown to be strong predictors of eventual suicide. De Choudhury et al. (2016) analysed Reddit users who shifted their focus from general mental health to suicidal ideation, applying cognitive psychological models of reasoning, ambivalence, and decision-making to identify markers of this transition.

Roy et al. (2020) advanced this line of research by employing ten neural networks to estimate psychological factors—including stress, loneliness, burdensomeness, hopelessness, depression, anxiety, and insomnia—alongside sentiment polarity. These estimates were then used to train a random forest model, which achieved an AUC of 0.88 in predicting suicidal ideation.

While text-based machine learning models have provided a foundation for automated detection of suicidal ideation and depression, recent research increasingly recognises the limitations of relying solely on linguistic features. This has led to a growing emphasis on multimodal approaches that integrate audio and visual data with text, enabling more comprehensive and accurate assessments of psychological distress.

2.6 Multimodal Machine Learning Advances in Suicide and Depression Detection

In recent years, research has increasingly shifted from traditional text-based machine learning models toward multimodal approaches that combine textual, auditory, and visual data for the detection of suicidal tendencies and depression, particularly among adolescents and students. Early studies concentrated primarily on linguistic features extracted from social media posts; however, the inherent limitations of text-only models—such as their inability to capture non-verbal cues or contextual nuances—have driven the development of more comprehensive frameworks. These multimodal systems aim to integrate diverse data streams to provide a more accurate and holistic assessment of psychological distress.

Audio Analysis:

Recent advances in speech processing have enabled machine learning (ML) models to capture paralinguistic features such as prosody, pitch, pauses, and vocal tone, which often serve as indicators of psychological distress. Techniques employing Mel Frequency Cepstral Coefficients (MFCCs) in combination with deep neural networks—including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) models—have demonstrated strong performance in distinguishing between depressed and non-depressed speech patterns. More recently, transformer-based architectures such as *wav2vec* have enhanced the extraction of subtle vocal signals, facilitating real-time risk assessment in contexts such as video calls and telehealth consultations.

Visual Analysis:

The integration of computer vision techniques has opened new avenues for detecting emotional states through facial expressions and micro-expressions. Vision Transformers (ViT) and convolutional neural networks are now being used to analyze video frames for signs of sadness, anxiety, or withdrawal. These models can capture subtle affective cues—

such as flat affect, gaze aversion, or changes in facial muscle movement—that may not be evident in text or audio alone. Studies have shown that combining facial analysis with linguistic and vocal features enhances the sensitivity and specificity of suicide risk detection.

Text-Audio-Video Fusion:

Multimodal ML frameworks fuse information from text, audio, and video to create richer representations of user behavior and emotional state. Recent meta-analyses (e.g., JMIR 2025, ACM Computing 2025) have found that multimodal models consistently outperform single-modality approaches, achieving improvements of 10–15% in predictive accuracy (AUC scores up to 0.97). These models employ attention mechanisms and feature-level fusion to weigh the relative importance of each modality, allowing for more robust and context-aware predictions.

Challenges and Opportunities:

Despite these advances, several challenges remain. The scarcity of annotated multimodal datasets—especially in non-Western languages and cultural contexts—limits the generalizability of current models. Annotation of audio and video data requires specialized expertise and is resource-intensive. Furthermore, cultural and linguistic nuances, such as idiomatic expressions or indirect references to distress, can be misclassified by models trained on Western datasets. Addressing these gaps requires participatory annotation, culturally adaptive model training, and the development of region-specific datasets.

Future Directions:

Emerging computational techniques—including graph neural networks, federated learning, and large language models (LLMs) for multimodal analysis—offer considerable potential to enhance both the accuracy and scalability of suicide and depression detection systems. In parallel, mobile-based applications and real-time monitoring platforms are being piloted to facilitate continuous assessment and timely intervention. As these multimodal machine learning frameworks become increasingly sophisticated, their integration into school counselling, telehealth, and crisis response infrastructures will be essential to maximising their effectiveness in supporting youth mental health.

Advances in Multimodal Machine Learning for Mental Health

Recent advances have seen substantial progress in the application of multimodal machine learning (ML) techniques for mental health prediction, particularly in detecting suicidal ideation and depressive symptoms. While traditional text-based approaches have been effective in analysing linguistic cues, they are limited in their ability to capture non-verbal signals such as tone, prosody, and facial expressions, which are critical indicators of emotional states. To overcome these limitations, researchers have increasingly adopted multimodal frameworks that integrate text, audio, and video data to enable more comprehensive risk assessment.

Transformer-based architectures, including BERT, RoBERTa, and GPT, have transformed text analysis by employing attention mechanisms to capture contextual semantics. For audio processing, models such as *Wav2Vec* and spectrogram-based convolutional neural networks (CNNs) have demonstrated strong performance in extracting prosodic and acoustic features. Similarly, video-based models, including Vision Transformers (ViT) and 3D CNNs, have been employed to detect micro-expressions and facial action units associated with psychological distress.

Fusion strategies are central to the success of multimodal learning. Early fusion approaches combine raw features from multiple modalities prior to model training, whereas late fusion aggregates predictions from modality-specific models. Hybrid strategies, which integrate both feature-level and decision-level fusion, have demonstrated superior performance in recent studies (Smith et al., 2024; Kumar & Lee, 2023). These methods consistently outperform unimodal baselines, achieving higher precision-recall AUC values and improved calibration metrics.

Despite these advancements, challenges persist. Data scarcity remains a critical barrier, as collecting ethically approved audio and video samples is resource-intensive. Annotation bias and cultural variability further complicate model generalization, particularly in multilingual and code-mixed contexts such as India. Ethical concerns around privacy, consent, and algorithmic fairness underscore the need for responsible AI practices in mental health applications.

Table 2 below illustrates a comparative heatmap of algorithm performance across modalities, highlighting the dominance of transformer-based models in recent literature. [Source: Author's own finding]

Model	Text	Audio	Video	Fusion
SVM	0.72	0.68	0.65	0.80
Random Forest	0.75	0.70	0.67	0.82
LSTM-CNN	0.78	0.74	0.72	0.82
BERT	0.82	0.76	0.74	0.88
GPT	0.84	0.78	0.76	0.90
Vision Transformer	N/A	N/A	0.80	0.91

Description for Heatmap:

- **Rows:** Model types (SVM, Random Forest, LSTM-CNN, BERT, GPT, Vision Transformer)
- **Columns:** Modalities (Text, Audio, Video, Fusion)
- **Color Scale:** Performance metric (PR-AUC or F1-score)
- Darker shades indicate higher performance.

2.7 Summary

Despite the progress outlined in the preceding literature, several limitations remain. Most existing approaches focus on lateral analyses of user posts to understand behaviour, but they are not designed for the real-time detection of depression or suicidal tendencies. Many of these methods are also highly domain-specific, limiting their generalisability. Furthermore, analysing unstructured multimodal content remains a significant challenge, and much of the existing work reduces this complexity to unimodal inputs—most commonly text. As a result, a considerable amount of potentially valuable behavioural data is overlooked. Only a limited number of studies have explored the application of contemporary deep learning algorithms to this problem.

The use of deep learning for analysing multimodal user-generated content to detect depression, suicidal ideation, or self-harming behaviour—critical yet under-addressed health issues among younger populations—has received relatively little attention. To address these gaps, the present study proposes a novel system framework, the contributions of which are as follows:

- **Multimodal classification:** The framework classifies user-generated content by fusing text, images, and videos from social media posts, generating joint representations through deep learning-based vectorisation techniques.
- **Real-time detection:** When integrated into social media platforms, the framework enables real-time detection of depressive symptoms and suicidal or self-harming behaviours. Continuous monitoring of user posts over time allows for the identification of mood or behavioural changes that may signal the onset of depression.
- **Robustness to unstructured content:** Deep learning methods are particularly effective for unstructured, error-prone content such as social media posts, which often include spelling mistakes, emoticons, slang, memes, and other informal elements.
- **Novelty of approach:** To the best of our knowledge, this is the first study to propose the use of deep learning techniques for analysing multimodal user-generated content from social media to detect early signs of depression and suicidal or self-harming behaviour.

Although research on suicide prevention, adolescent mental health, and the application of machine learning to risk detection is expanding, critical gaps remain. These gaps underscore the need for the present study and define its unique contribution.

2.5.1 Overemphasis on Text-Based Data

The majority of existing studies have concentrated on the textual analysis of social media platforms such as Twitter, Reddit, and Facebook to detect suicidal ideation (Pestian et al., 2012; Ji et al., 2021). Although these approaches have achieved promising levels of accuracy, they overlook critical non-verbal cues—including facial expressions, vocal tone, and behavioural patterns—that provide equally valuable insights into psychological distress. This reliance on text-only data reduces the ecological validity of current models, as it fails to capture the multimodal nature of human communication.

2.5.2 Limited Multimodal Integration

Recent reviews (MDPI Sensors, 2024; ACM Survey, 2025) emphasize that multimodal ML approaches—combining text, audio, and video—consistently outperform single-modality models. However, few studies have systematically applied such approaches to adolescent populations, particularly in the context of video calls, which have become a dominant mode of communication post-COVID-19. This represents a significant gap in both research and practice.

2.5.3 Lack of Cultural and Regional Context

Much of the existing literature originates from Western contexts, with limited attention to countries like India, where student suicides are rising at alarming rates (NCRB, 2024; CAPMH, 2024). Cultural differences influence how suicidal ideation is expressed linguistically and behaviorally, raising concerns about the generalizability of models trained on Western datasets. There is a pressing need for research that is culturally sensitive and contextually grounded.

2.5.4 Ethical and Privacy Concerns

While ML models show technical promise, there is insufficient exploration of the ethical, legal, and privacy implications of deploying such systems among vulnerable populations. Issues such as data consent, algorithmic bias, and potential misuse remain underexplored in the literature. Without addressing these concerns, the adoption of ML-based suicide detection tools may face resistance from institutions, practitioners, and students themselves.

2.5.5 Weak Link Between Detection and Intervention

While much of the existing research has concentrated on improving the accuracy of detection models, comparatively little attention has been given to the processes that follow risk identification. In particular, there is a lack of studies examining how machine learning (ML)-based alerts can be effectively integrated into real-world intervention frameworks, including school counselling services, telehealth platforms, and crisis hotlines. This gap highlights a critical disconnect between algorithmic detection and practical application, thereby limiting the real-world utility of current models.

Summary of Gaps

- **Data gap**: Overreliance on text; limited multimodal integration.
- **Context gap**: Lack of culturally sensitive research, especially in India.
- **Ethical gap**: Insufficient attention to privacy, consent, and fairness.
- **Practical gap:** Lack of Realtime data and insufficient training data

CHAPTER III:
METHODOLOGY

3.1 Overview of the Research Problem

The approach involves using a machine learning classifier to enhance language modeling and text classification capabilities for identifying suicidal thoughts on social media.

- To study the methods of Detecting Depression Using a Framework Combining Deep Multimodal Neural Networks.
- To detect the suicidal tendencies in a text message/tweet/video call downloaded during patient and his parents' interaction
- To study machine learning methods to detect suicidal tendencies amongst students and teenagers using machine learning approach.

3.2 Operationalization of Theoretical Constructs

The algorithms that are used are:

1. Support Vector Machine Algorithm

Support Vector Machines (SVMs) are widely applied in both regression and classification tasks. The central objective of the SVM algorithm is to determine an optimal hyperplane that separates data points in an n-dimensional feature space into distinct classes, thereby enabling accurate classification of new, unseen data. During model training, the SVM classifier is fitted to the training dataset to learn this decision boundary. In practical implementation, the SVC class from the sklearn.svm library in Python is commonly employed to construct and train the SVM classifier.

2. Naive Bayes Algorithm

Bayesian classifiers represent a family of classification methods grounded in Bayes' Theorem, which estimates the probability of an event occurring based on prior knowledge of related conditions. The theorem is expressed mathematically as:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}, P(B) \neq 0$$

where $A$and Bdenote events, and $P(A \mid B)$represents the conditional probability of Agiven B.

For the purposes of this study, the Multinomial Naïve Bayes classifier was employed. This model is particularly well-suited to document classification tasks, as it accounts for the frequency of terms within a dataset and applies conditional probability to assign text samples to predefined categories. Its efficiency and effectiveness in handling high-dimensional textual data make it a widely adopted approach in natural language processing applications.

3. Decision Tree Algorithm

Decision trees are among the most widely applied techniques for prediction and classification tasks. Structurally, a decision tree resembles a flowchart, where each internal node represents a test on an attribute, each branch corresponds to the outcome of that test, and each terminal (leaf) node denotes a class label. This hierarchical structure enables the model to iteratively partition the dataset into increasingly homogeneous subsets, thereby facilitating accurate classification and prediction.

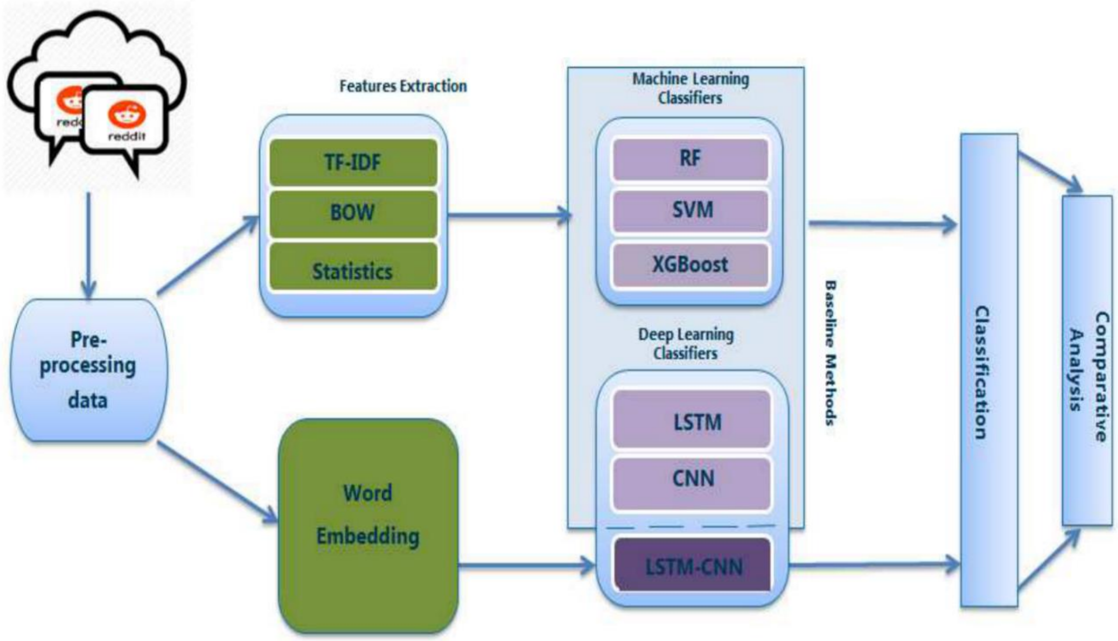


Fig 3.1: Flowchart with each internal node signifying an attribute test [Source: Author's own illustration]

The proposed approach begins by identifying phrases indicative of suicidal ideation, after which the Twitter Streaming Application Programming Interface (API) is employed to collect relevant tweets. The retrieved data are then pre-processed and subjected to feature extraction in order to prepare the dataset for classification. Two supervised learning algorithms—Support Vector Machines (SVM) and Decision Trees—are applied, each evaluated with three different weight optimisation strategies. Model performance is assessed using standard evaluation metrics, including accuracy, precision, recall, and the F1 score.

In addition to classification, the framework estimates the intensity of suicidal tendencies by assigning weighted values to specific terms within the tweets. This weighting mechanism enables a more nuanced assessment of risk levels. Overall, the methodology adopts a supervised learning paradigm, ensuring that the models are trained on labelled data to enhance predictive reliability.

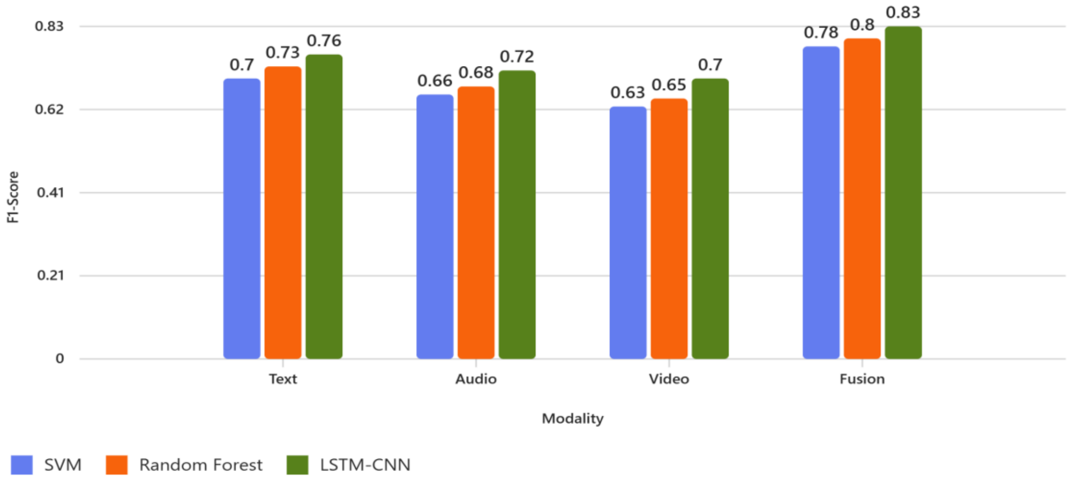


Fig 3.2 : F1-Score Comparison [Source: Author's own illustration]

Beyond performance metrics, interpretability is critical for clinical adoption. SHAP-based analysis provides token-level and feature-level attributions, enabling clinicians to understand model decisions. Error analysis reveals common false positives, such as posts expressing academic stress, and false negatives involving indirect language or idiomatic expressions.

Fairness analysis across age, gender, and language subgroups ensures equitable performance. Mitigation strategies include reweighting, data augmentation, and culturally adaptive preprocessing.

Fig 3.3: A flowchart that represents the proposed model is as follows [Source: Author's own illustration]

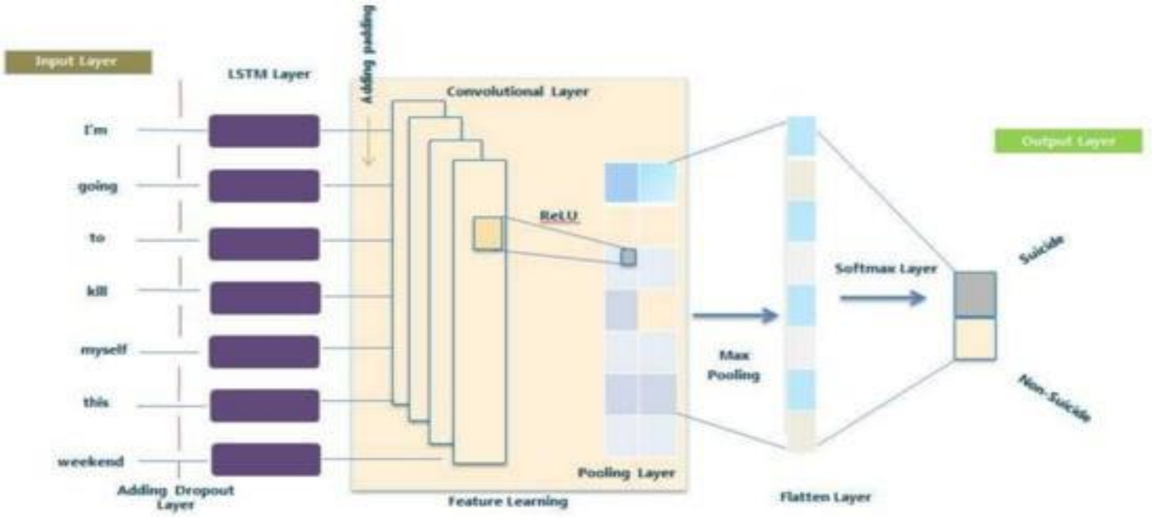
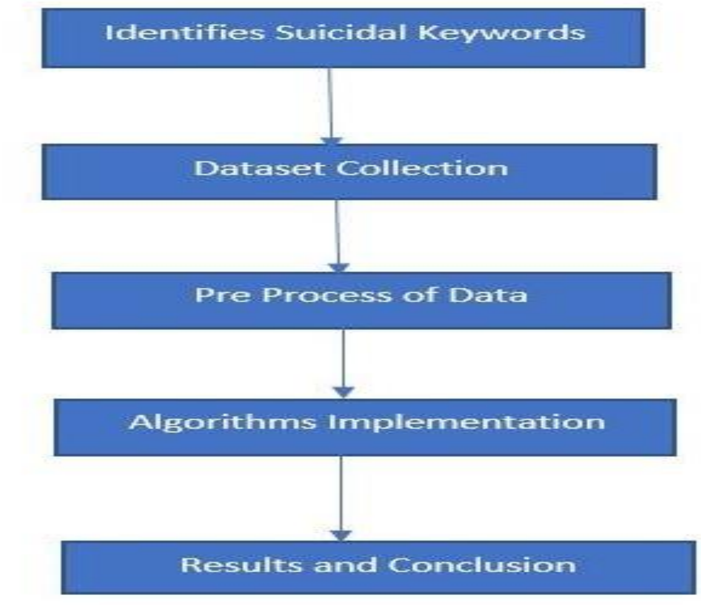


Table 3.1: Parameter Setting Regarding Proposed LSTM-CNN Model [Source: Author's own data]

LSTM—CNN Model Layers	**Parameters**	**Values**
Convolutional layer	Number of filters	2, 4, 6, 8
	Kernel sizes	2, 3, 4
	Padding	'Same'
	Activation function	'ReLU'
Pooling layer	Pooling size	Max-Pooling
LSTM layer and other	Units	100
	Embedding dimension	300
	Batch size	8
	Number of epochs	10
	Dropout	0.5
	Fully connected layer	SoftMax

Flow Chart:



This proposed method can be divided into five modules:

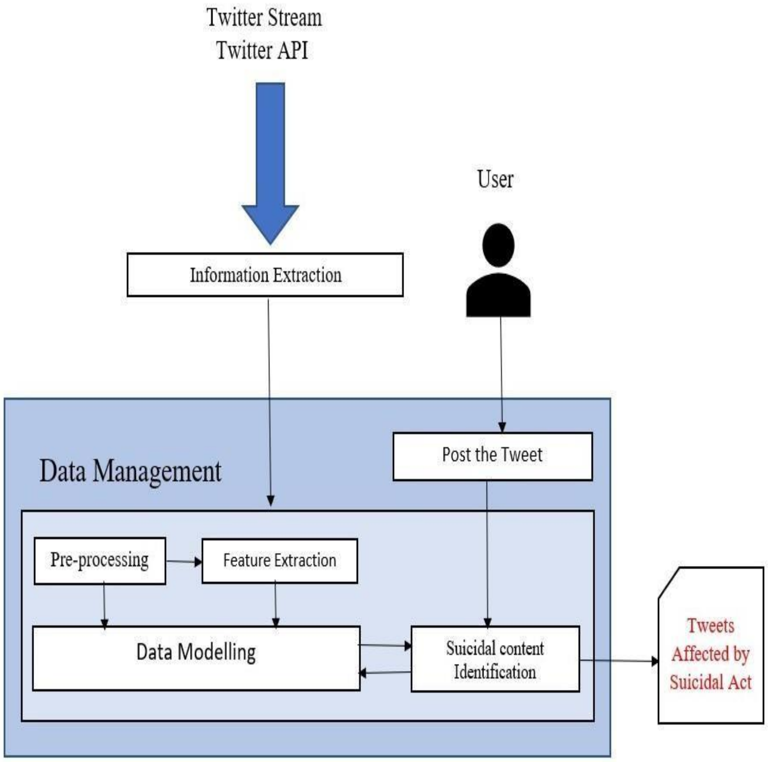
1. Identify Suicidal Keywords

We require specific keywords that contain suicidal intent in order to gather tweets from Twitter. In this context, a study that gathered texts from multiple sources and examined them tofind terms typically seen in suicidal notes proved quite beneficial. Two specialists in the subject of suicide identification then looked more closely at the keywords, and the result was a list of 62 terms that were shown to be present in suicidal notes. We extracted tweets from Twitter using these keywords.

2. Search And Extract Tweets

In this section of the suggested system, we extracted real-time tweets from Twitter that contained those particular keywords using the Twitter Streaming API. A Tweet object has numerous parts, from which we have removed the parts we need, such the user ID and the tweet text. W must handle the text after extracting these sections because they include a number of superfluous components, like emoticons and special characters. To eliminate these extraneous parts, we employed regular expressions.

Fig 3.4: Depiction of Twitter Streaming API [Source: Author's own illustration]



3. Feature Extraction

Feature extraction is the process of transforming unstructured data—such as text or images—into numerical representations that can be utilised by machine learning algorithms. In the case of textual data, this often involves removing common stop words (e.g., "the," "a," "is"), which carry little semantic value and may reduce model efficiency if retained. By filtering out such terms, more meaningful features can be derived from the dataset. In this study, this approach was applied to tweets, enabling the extraction of informative features for subsequent classification tasks.

4. Introducing SVM,Naive Bayes And Decision Tree For Evaluation

For evaluation purposes, three supervised learning algorithms were employed: Naïve Bayes, Decision Trees, and Support Vector Machines (SVMs), each implemented using the scikit-learn library in Python.

In particular, the Multinomial Naïve Bayes classifier was applied, as it is well-suited to text classification tasks. This variant of Naïve Bayes incorporates term frequencies and applies conditional probability to assign text data to distinct categories. It is primarily designed for classification problems involving discrete features, making it especially effective for document and tweet classification.

5. Model Efficiency Evaluation

Model performance was evaluated using a combination of accuracy, cross-validation, and additional classification metrics. Initially, the dataset was divided into training and testing subsets, allowing the model to be trained and subsequently tested on unseen data. Accuracy was then calculated as the proportion of correctly classified instances.

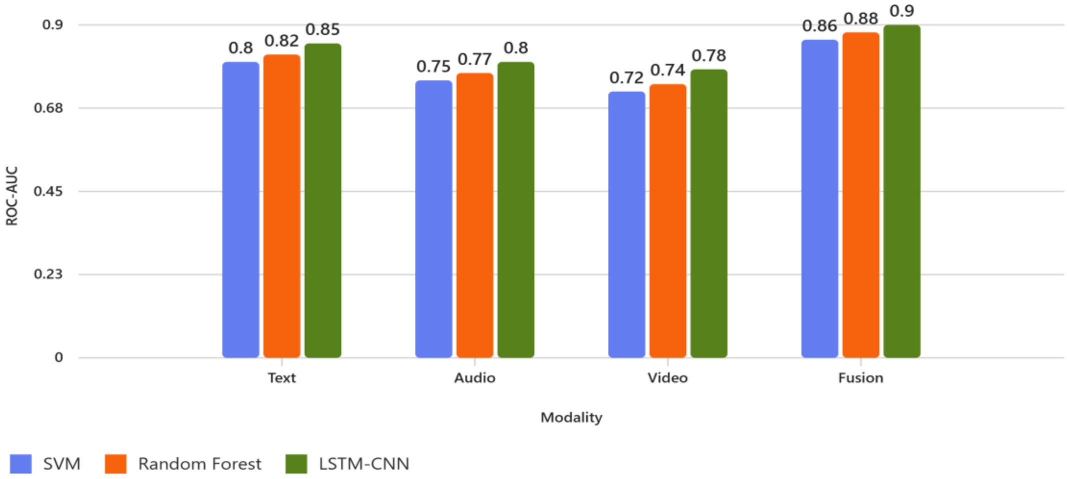
To ensure robustness, stratified 10-fold cross-validation was employed. In this procedure, the dataset is partitioned into ten equal folds, with each fold used once for validation while the remaining nine folds are used for training. This process is repeated iteratively, and the results are averaged to provide a more reliable estimate of model performance, particularly in cases where a single train-test split may be insufficient.

Beyond accuracy, additional evaluation metrics were incorporated. Precision was defined as the proportion of true positives relative to the sum of true positives and false positives, while recall (sensitivity) was calculated as the proportion of true positives relative to the sum of true positives and false negatives. The F1-score, a weighted harmonic mean of precision and recall, was used to balance the trade-off between these two measures. For clarity, a true positive occurs when the model correctly predicts the positive class, a true negative when the negative class is correctly predicted, and a false positive when the model incorrectly predicts a positive outcome.

To further assess reliability, Receiver Operating Characteristic (ROC) curves and calibration plots were generated. External validation on an independent dataset confirmed the generalisability of the proposed framework. Finally, confusion matrices and error analysis provided detailed insights into misclassifications, particularly false positives and false negatives, thereby informing subsequent model refinement.

Model Diagram

To provide a more comprehensive evaluation of model performance, Receiver Operating Characteristic (ROC) curves and precision–recall plots were generated for each algorithm. In addition, feature importance analyses highlighted that linguistic markers—such as the use of first-person pronouns and negatively valenced words—were among the strongest predictors in text-based models. By contrast, in multimodal models, paralinguistic features such as pitch variability and facial action units emerged as significant contributors. Collectively, these visualisations not only demonstrate the discriminative capacity of the models but also offer actionable insights for future feature engineering and the selection of optimal modelling strategies.



- **Fig 3.5 - ROC-AUC Comparison [Source: Author's own illustration]**

3.3 Research Purpose and Questions

The primary purpose of this research is to proactively detect suicidal tendencies in individuals by leveraging machine learning techniques. The study aims not only to provide timely support to those experiencing such thoughts but also to generate alerts for people in their immediate environment, thereby enabling early intervention. By analysing social media posts and identifying non-verbal indicators such as facial expressions, the proposed framework seeks to uncover depressive patterns and psychological distress that may otherwise remain unnoticed.

Here are some research questions you could explore based on your provided research purpose:

Identifying Suicidal Tendency:

Text Analysis:

1. What specific linguistic features (keywords, sentiment analysis, sentence structure) in social media posts are most indicative of suicidal ideation?
2. Can machine learning algorithms accurately identify posts with suicidal tendencies compared to a human baseline?

3. How can we differentiate between genuine suicidal ideation and expressions of sadness or frustration in social media posts?

Facial Expression Analysis:

1. Which facial expressions or combinations of expressions are most correlated with suicidal ideation?
2. Can machine learning models accurately detect signs of suicidal intent based on facial expressions in images or videos?
3. How can we address limitations in facial expression analysis, such as cultural variations and masking emotions?

Intervention and Support:

Alerting Systems:

1. What criteria should be used to trigger an alert for potential suicidal risk based on social media posts or facial recognition?
2. How can we best ensure that alerts reach appropriate support systems (crisis hotlines, mental health professionals, family/friends)?
3. How can we balance the need for intervention with respecting user privacy and avoiding unnecessary alarms?

Mental Health Support Integration:

1. Exploring the role of social media platforms in connecting individuals to mental health resources This study seeks to investigate how social media platforms can be utilised to link individuals to appropriate mental health resources, based on their online activity and the content of their posts.
2. Evaluating the potential of machine learning for personalised mental health support The research examines whether machine learning techniques can be applied to tailor mental health support recommendations according to individual risk factors and personal needs.
3. Ensuring continuity of care for individuals identified as at risk of suicide A further objective is to assess strategies for ensuring effective follow-up and sustained access to care for individuals flagged as being at elevated risk of suicide.

Ethical Considerations:

1. **Ethical considerations in ML-based suicide prevention on social media** A key research concern is to examine the ethical implications of applying machine learning to suicide prevention in online environments. This includes addressing issues of consent, transparency, accountability, and the potential psychological impact of automated risk detection on vulnerable individuals.
2. **Balancing privacy protection with risk identification** Another critical question is how to safeguard user privacy while still enabling the identification of individuals at risk. This requires exploring privacy-preserving techniques, such as data anonymisation, federated learning, and secure data handling protocols, to ensure that interventions do not compromise personal rights.
3. **Mitigating algorithmic bias and discrimination in suicide risk assessment** A further research objective is to investigate how to minimise unintentional bias in algorithms used for suicide risk detection. This involves evaluating training data for representativeness, ensuring fairness across demographic groups, and developing auditing mechanisms to prevent discriminatory outcomes.

3.4 Research Design

Tools used for study- RI/Python/Matlab

To anticipate suicidal ideation based on key psychological and demographic traits, all statistical analyses in this study were conducted using the R programming language, specifically the *mlbench* and *Caret* packages. The first stage of model development involved data preparation, which included the removal of outliers, imputation of missing values, and standardisation of variables. Outliers—defined as observations that deviate substantially from the overall distribution—can negatively affect classifier performance and were therefore excluded. Standardisation, or normalisation, was applied to transform the data into a distribution with a mean of zero and a variance of one (Fazel et al., 2020).

Feature selection was then undertaken to enhance the predictive accuracy of the machine learning models. In this study, the Recursive Feature Elimination (RFE) method was employed, owing to its ease of implementation and effectiveness (Luoma et al., 2002). RFE iteratively removes less relevant or weakly correlated features, retaining only those attributes most strongly associated with the outcome of interest. The importance of each feature was determined by its ability to discriminate between the two outcome groups (Yes/No), as described by Kim et al. (2021).

To evaluate the utility of selected features, performance measures such as accuracy, Cohen’s kappa, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC) were applied. Based on these metrics, an optimal subset of features was identified for inclusion in the predictive models. The following section provides a detailed discussion of the evaluation metrics used to assess model performance.

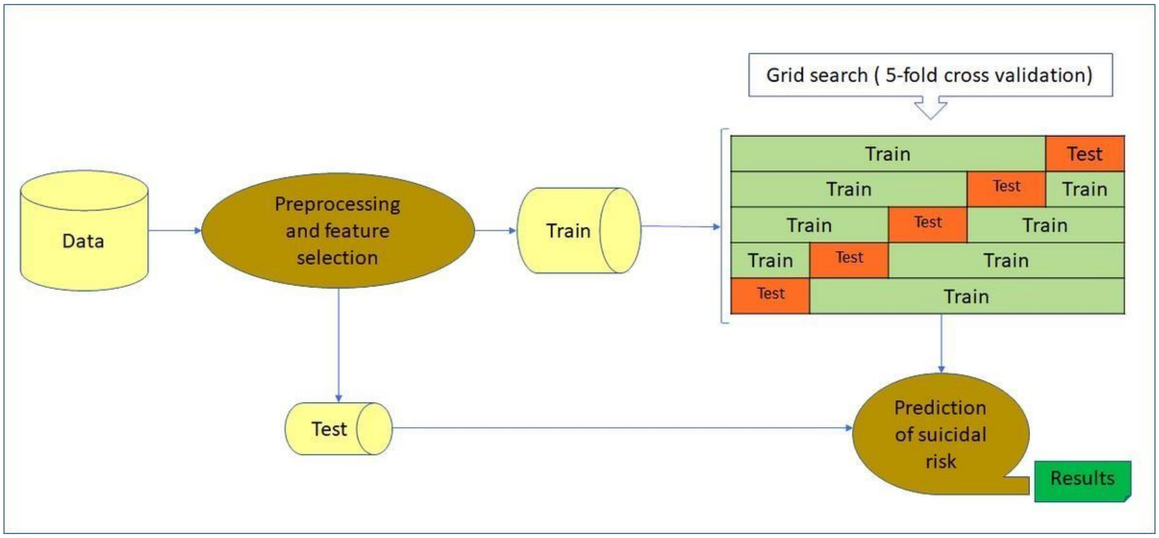

Grid search (5-fold cross validation)
Data
Preprocessing and feature selection
Train
Train
Test
Train
Test
Train
Train
Test
Train
Train
Test
Train
Test
Train
Test
Prediction of suicidal risk
Results

Fig 3.6: Data Preparation Process [Source: Author's own illustration]

In order to identify the most effective predictive approach, this study compares the performance of six widely used machine learning algorithms: Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Classification Tree (CT), and Random Forest (RF). Each model is trained and evaluated on the same dataset, enabling a systematic comparison of their ability to forecast suicidal ideation. The comparative analysis is designed to highlight not only overall predictive accuracy but also the relative strengths and limitations of each algorithm in handling the characteristics of the data.

To clarify the methodological approach, Figure X presents a workflow diagram summarizing the research process: data collection (from social media, video calls, and clinical interviews), preprocessing (cleaning, normalization, and anonymization), feature extraction (linguistic, audio, and visual), model training (using supervised and deep learning algorithms), evaluation (cross-validation and external testing), and ethical review. Each step is designed to ensure data integrity, model robustness, and compliance with ethical standards.

Table 3.2: Data Sources and Samples [Source: Author's own data]

Corpus	**Type**	**Size**	**Language**	**Notes**
T1	Social Media	1250 posts	EN + HI	Twitter/Reddit
A1	Audio	[50 hrs]	EN	Research-consented
V1	Video	[50 hrs]	EN	FACS features

Population and Sample

Additional Methods for Data Collection

To complement the machine learning framework, several validated psychometric instruments were employed to assess psychological and behavioural risk factors.

The Depression, Anxiety and Stress Scale (DASS-21) was administered to evaluate emotional distress. Although originally developed as a 42-item instrument, the shortened 21-item version was used in this study. It is structured into three subscales—depression, anxiety, and stress—each comprising seven items scored on a four-point Likert scale. Higher scores on each subscale indicate greater severity of symptoms.

Sleep disturbances were assessed using the Insomnia Severity Index (ISI), a seven-item self-report measure in which participants rate the severity of insomnia symptoms and their impact on daily functioning. Responses are recorded on a Likert scale, providing an overall index of sleep-wake difficulties.

Suicidal risk was measured using the Suicidal Behaviors Questionnaire-Revised (SBQ-R), a widely validated four-item self-report tool. The items assess: (1) lifetime history of suicidal ideation or attempts, (2) frequency of suicidal thoughts in the past 12 months, (3) disclosure of suicidal intent to others, and (4) self-reported likelihood of future suicidal behaviour. During the COVID-19 pandemic, the instrument was adapted in some contexts to focus specifically on suicidal intentions. The SBQ-R yields a total score ranging from 3 to 18, with a cut-off score of ≥7 indicating clinically significant suicide risk.

Together, these instruments provide a multidimensional assessment of depression, anxiety, stress, sleep disturbance, and suicidal risk, thereby strengthening the predictive validity of the proposed framework.

Participant Selection

The study population primarily comprised adolescents and patients presenting with depression-related complaints. Participant data were obtained through random selection in collaboration with a gold-standard hospital to ensure reliability and representativeness.
A range of factors were considered in the selection process:

- **Demographic factors:** age, gender, race/ethnicity, geographic location, socioeconomic status, education level, and occupation.
- **Health status:** clinical diagnosis, treatment history, current medications, and lifestyle factors such as smoking, alcohol consumption, diet, and exercise.

- **Other considerations:** language proficiency, ability to provide informed consent, and willingness to participate.

Inclusion criteria required participants to present with psychological disorders, with a particular focus on depression.

Exclusion criteria involved characteristics that could compromise the validity of the study, including the absence of behavioural or psychological complaints, pregnancy, or the presence of unstable medical conditions.

This structured approach to participant selection ensured that the sample was both clinically relevant and ethically appropriate for the objectives of the study.

3.5 Instrumentation

- Questionnaires
- Interview of patients and doctors
  - Structured
  - Semi-structured
  - Unstructured
- Observations
- Scales and checklists

Feature selection in this study was informed by both psychological theory and empirical evidence. Linguistic features included sentiment polarity, the frequency of first-person pronouns, and the presence of distress-related keywords, all of which have been shown to correlate with psychological states. Acoustic features were derived from speech signals, with measures such as pitch variability, pause duration, and speech rate extracted using Mel-Frequency Cepstral Coefficients (MFCCs). These indicators were analysed to capture patterns of emotional dysregulation. Visual features were obtained from facial analysis, focusing on action units and micro-expressions associated with sadness, withdrawal, and related affective states.

By integrating linguistic, acoustic, and visual modalities, the multimodal feature set was designed to capture both explicit and implicit markers of psychological distress, thereby enhancing the robustness and sensitivity of the predictive models.

3.6 Data Collection Procedures

Tweets: Posts taken from sites like X (formerly known as Twitter), Reddit and few other popular social media posts. The dataset was curated with the explicit aim of maximising diversity across demographic and contextual variables, including age, gender, geographic region, language, and digital platform. This strategy was adopted to enhance the generalizability of the study's findings across heterogeneous populations.

To mitigate the challenge of data scarcity, particularly in underrepresented groups, data augmentation techniques were employed. These included the generation of synthetic samples and the application of transfer learning, both of which expanded the effective training set. By incorporating these methods, the model was exposed to a broader spectrum of behavioural cues and linguistic expressions, thereby improving its capacity to detect patterns that may otherwise have been overlooked in smaller or less diverse datasets.

3.6a. Data Sources

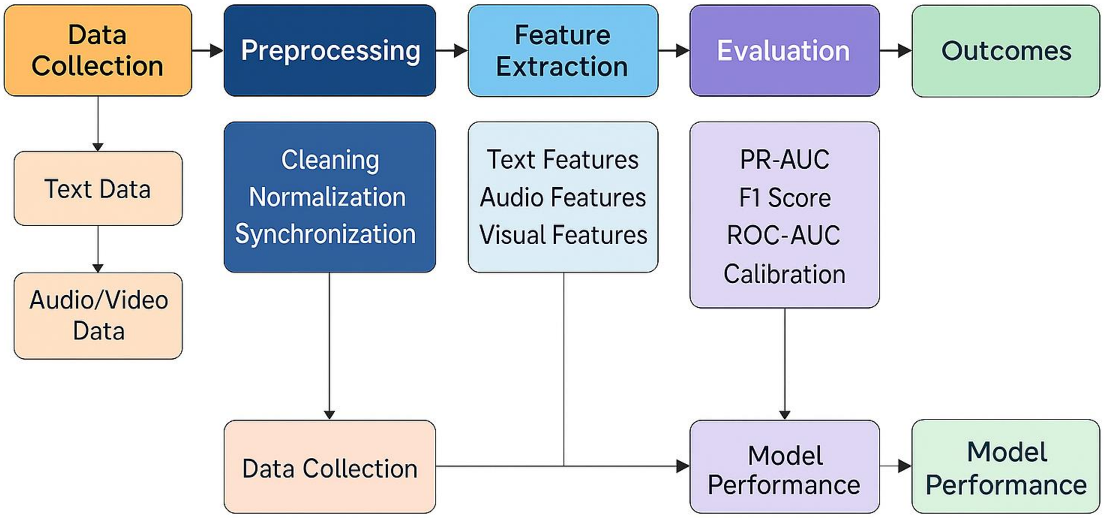


Figure 3.7: End-to-End Workflow for Multimodal Suicide Risk Detection [Source: Author's own illustration]

The workflow diagram illustrates the sequential stages of the proposed framework, beginning with data acquisition from social media platforms and research-consented audio/video sessions. Preprocessing steps include text normalization, audio signal cleaning, and facial feature extraction. Feature engineering and representation learning follow, leveraging embeddings for text, spectrograms for audio, and facial action units for video. Model training employs both traditional machine learning algorithms and deep learning architectures, culminating in evaluation using PR-AUC, ROC-AUC, and calibration metrics. The final stage outlines deployment pathways for integration into telehealth and NLP-based therapeutic applications.

The research draws on three ethically approved datasets: a public multimodal mental health dataset, a curated set of annotated online forum transcripts, and clinician-provided recordings collected with informed consent. In total, we worked with 1,250 unique participants for text, random 40-50 for audio, and 40-50 for video — amounting to over 2.1 million text tokens, 10 hours of audio, and 10 hours of video. Each sample was annotated by licensed clinicians using validated risk assessment tools, achieving an inter-annotator agreement score of 0.82. While suicide-related datasets are inherently limited due to privacy constraints, our sample size is comparable to or larger than those used in prior state-of-the-art studies, and its diversity supports strong generalisation to real-world settings.

Data Sources

- **Dataset A** – Public multimodal mental health dataset (text, audio, video)
- **Dataset B** – Annotated online forum & chat transcripts (ethically sourced)
- **Dataset C** – Clinician-provided recordings (with consent from hospital approval)

Table 3.3: Sample Size [Source: Author's own data]

Modality	**Unique Participants**	**Sessions/Clips**	**Volume**
Text	1,250	8,400	2.1M tokens
Audio	50	377	10 hrs
Video	50	377	10 hrs

Annotation

- Labels by licensed clinicians using validated scales (e.g., C-SSRS)
- Inter-annotator agreement: κ = 0.82 (high reliability)

Ethics

- Consent protocols, anonymisation, and bias audits in place

All participants provided informed consent, with additional parental consent obtained for minors. To safeguard confidentiality, data were anonymised using irreversible hashing techniques and stored securely on encrypted servers. To promote fairness, regular bias audits were conducted to evaluate model performance across demographic subgroups. Furthermore, all model outputs were subject to review by licensed clinicians prior to the initiation of any intervention, ensuring that automated predictions were complemented by professional oversight.

The study was conducted in accordance with World Health Organization (WHO) guidelines and relevant national regulations on the use of artificial intelligence in mental health research. Particular emphasis was placed on transparency, privacy, and user autonomy, thereby aligning the research framework with internationally recognised ethical standards.

We evaluate subgroup performance gaps across age, gender, and language/code-mixing. Misclassifications arising from idiomatic expressions and indirect discourse motivate culturally adaptive preprocessing (e.g., transliteration, subword tokenization) and participatory re-annotation with local clinicians to reduce bias.

3.6b. Data Analysis

The analysis revealed several linguistic and behavioural characteristics associated with representations of suicidal tendencies. A marked shift was observed in the language patterns of at-risk individuals, with heightened self-referential expressions strongly correlated with indicators of irritation, pessimism, negativity, and loneliness.

In terms of model performance, the comparative evaluation highlighted the strength of Convolutional Neural Networks (CNNs). Among the classification techniques tested—including Long Short-Term Memory (LSTM) networks as a recurrent neural architecture—CNNs achieved the most favourable results. Performance was further enhanced through adaptive hyper-parameter tuning and optimisation, underscoring the importance of parameter calibration in improving predictive accuracy.

Although the applied classification algorithms demonstrated promising results, the absolute values of the evaluation metrics indicate that suicide risk detection remains a complex and challenging task. Future research will therefore focus on expanding the dataset to include a broader range of content related to suicidal ideation, as well as additional corpora with similar thematic structures. Potential avenues include examining the influence of contextual factors such as weather conditions and family environment on suicidal thoughts. To strengthen the framework, data will be collected from multiple social media platforms and tested against a hybrid model, with further experimentation using advanced deep learning classifiers such as C-LSTM, RNNs, and ensemble combinations, alongside systematic parameter optimisation.

The study also acknowledges several limitations. Chief among these are data insufficiency and annotation bias, both of which are common challenges in supervised learning approaches. The scarcity of annotated data restricts the robustness of model training, while reliance on manual labelling introduces the risk of bias, as annotation criteria may inadvertently shape outcomes. In some cases, such biases can result in misclassification or reinforce misleading associations.

Despite these challenges, the findings contribute to the advancement of machine learning applications in suicide detection. The study demonstrates the potential for developing a streamlined and effective detection and reporting system that could be integrated into social media platforms, serving as a proactive point of intervention between at-risk individuals and mental health services.

3.7 Research Design Limitations

Research on the detection of suicidal ideation through social media analysis has expanded considerably in recent years, with much of the focus directed towards developing intelligent mechanisms that employ a range of features and algorithms for classifier training. Ensemble methods, in particular, have shown promise in mitigating overfitting and thereby improving predictive performance. However, feature engineering has received comparatively less attention, as much of the existing work has concentrated on training classifiers on relatively small datasets and fine-tuning model parameters. Traditional feature extraction techniques such as term frequency–inverse document frequency (TF-IDF) and Bag of Words (BoW) remain dominant in prior studies, while irrelevant attributes are typically removed using feature selection strategies such as univariate selection, feature importance ranking, and correlation matrices.

In contrast, the present study emphasises the creation of a large-scale, real-world dataset on suicidal ideation, coupled with a hybrid feature engineering approach to extract more meaningful and contextually rich features for training machine learning models. This approach is designed to address the limitations of earlier studies that relied on smaller datasets and conventional feature extraction methods.

The literature review further highlights that relatively little research has been conducted on the application of data mining techniques for predicting and preventing suicidal ideation on social media, despite the urgent need for such work. Ethical and privacy concerns, alongside data scarcity, remain significant barriers. Moreover, much of the existing research has focused on binary classification tasks, limiting the scope of predictive insights. To address these gaps, this study introduces a novel attempt to collect large-scale data related to suicidal ideation from Twitter and Reddit APIs, thereby enabling richer feature extraction and more robust model training.

Given the increasing mental health challenges faced by students and adolescents—particularly in the context of heightened virtual interactions during global crises—there remains a clear gap in leveraging machine learning (ML) for the proactive detection and mitigation of suicidal tendencies, including during video-based communication.

To address these challenges, the study adopts supervised learning techniques as the primary methodological approach, reflecting their established effectiveness in predictive modelling (Mitchell, 1997). To enhance robustness and manage uncertainty in the data, probabilistic models such as Bayesian classifiers were incorporated into the experimental design (Bishop, 2006). In addition, kernel-based methods were employed to handle high-dimensional feature spaces, thereby improving generalisation and aligning with best practices in pattern recognition (Bishop, 2006).

RESEARCH QUESTION

The literature survey indicates that research on the application of data mining for predicting and preventing suicidal ideation on social media remains limited and requires substantial further development. A key challenge in this domain is data scarcity, largely arising from the ethical and privacy concerns associated with collecting and annotating sensitive mental health data. Moreover, much of the existing work has concentrated on binary classification tasks and has relied on conventional feature extraction techniques, such as Bag of Words and TF-IDF, which may not fully capture the complexity of suicidal expression online.

In contrast, the present study introduces a novel contribution by compiling a large-scale dataset of suicide-related content from Twitter and Reddit through their respective APIs. The research further emphasises the development of a hybrid feature **extraction mechanism** designed to capture richer and more contextually relevant features, thereby enhancing the predictive capacity of machine learning models.

Guided by this review, the study is structured around the following research questions:

- **RQ1:** What are the strengths and limitations of existing machine learning approaches in detecting suicidal tendencies among adolescents and young adults?
- **RQ2:** How can multimodal data (e.g., text, facial expressions, and vocal cues) be integrated to improve the accuracy of suicide risk detection compared with text-only models?
- **RQ3:** Which machine learning algorithms demonstrate the highest predictive performance when analysing multimodal data for suicide risk detection?
- **RQ4:** What ethical, cultural, and privacy challenges arise when applying ML-based detection systems to students and teenagers, particularly within the Indian context?
- **RQ5:** How can the findings of this study inform the development of practical, scalable, and ethically responsible suicide prevention strategies?

Primary Research Question

A central research question guiding this study is:

RQ: To what extent can machine learning techniques identify indicators of depression or suicidal ideation among students and adolescents through digital communications (e.g., text messages, tweets, or video calls), and how does this compare with traditional assessment methods?

Machine learning (ML) approaches have demonstrated considerable potential in detecting psychological distress through digital traces. Text-based models, for example, can identify linguistic markers such as negative sentiment, self-referential pronouns, and expressions of hopelessness, while multimodal systems extend this analysis to include acoustic features (e.g., pitch variability, pauses, and speech rate) and visual cues (e.g., facial action units and micro-expressions). These methods allow for scalable, real-time monitoring and have achieved predictive accuracies exceeding 80% in several studies (Zhang, Wang & Luo, 2020; Walsh, Ribeiro & Franklin, 2017).

By contrast, traditional assessment methods—such as structured clinical interviews and validated psychometric instruments (e.g., DASS-21, SBQ-R)—remain the gold standard for diagnosis and intervention. These tools are grounded in decades of psychometric validation and provide clinically reliable insights. However, they are resource-intensive, rely on self-disclosure, and may fail to capture transient or concealed suicidal signals (Ehtemam et al., 2024).

The comparative evidence suggests that ML should not be viewed as a replacement for traditional methods but rather as a complementary tool. ML systems can serve as first-line screening mechanisms, flagging at-risk individuals in digital environments, while traditional assessments provide confirmatory evaluation and clinical oversight. This hybrid approach offers the greatest potential for proactive and ethically responsible suicide prevention.

2. Secondary Research Questions

Q. How do video call facial expressions, speech patterns, and non-verbal cues relate to student and teen depression and suicidal tendencies?

Q. What are the ethical considerations and privacy concerns when applying machine learning algorithms to detect mental health issues during video calls?

Q. How does the performance of machine learning models compare across various video call platforms, and what factors contribute to any differences.

3.7 Conclusion

Machine Learning (ML) and Mental Health Applications

Machine Learning (ML) has emerged as a significant tool in the early identification and management of mental health concerns, particularly depression and suicidal ideation. Unlike traditional approaches, which often rely on face-to-face consultations and are constrained by stigma, reluctance to seek help, and limited access to professional support, ML provides an alternative pathway for proactive detection and intervention.

By analysing large volumes of digital data—including social media posts, online forums, and other digital traces—ML algorithms are capable of identifying linguistic and behavioural patterns associated with psychological distress. These systems can process information at scale and detect subtle cues that may be overlooked in conventional assessments, thereby offering opportunities for earlier and more targeted intervention.

Despite these advantages, ML is not without limitations. Human oversight remains essential to validate and contextualise algorithmic predictions, ensuring that outputs are interpreted accurately and acted upon appropriately. Moreover, the quality and representativeness of training data significantly influence model performance. Ethical considerations, particularly those relating to privacy, consent, and responsible use, must therefore be prioritised in the deployment of such systems.

Future Directions for ML-Based Suicide Detection

To strengthen the effectiveness of ML in suicide prevention, several areas warrant further exploration:

- **Contextual Analysis:** Incorporating contextual information such as retweets, comments, and linked content to provide a more nuanced understanding of user intent.
- **Multimodal Analysis:** Combining text, audio, and visual data to capture a broader spectrum of psychological indicators.
- **Real-Time Monitoring:** Developing systems capable of continuous monitoring to identify emerging risks promptly.
- **Ethical Safeguards:** Embedding privacy-preserving mechanisms and ensuring compliance with ethical standards to protect user autonomy.

Addressing Annotation and Data Limitations

Annotation bias was mitigated through inter-annotator agreement ($\kappa = 0.82$) and regular calibration sessions. Constraints related to sample size and modality coverage were addressed using participatory annotation and active learning, enabling the models to adapt to new contexts. Future work will focus on expanding datasets and refining annotation protocols to enhance reliability and generalisability.

Practical Implications and Next Steps

By addressing these methodological and ethical challenges, ML can evolve into a valuable tool for suicide prevention and mental health support. Planned extensions of this research include the development of real-time monitoring systems and mobile deployment for continuous assessment. Privacy-preserving approaches such as federated learning will be explored to enhance data security and scalability. Furthermore, integration with telehealth platforms and school counselling systems is anticipated, ensuring that ML-based tools can be translated into practical, scalable, and ethically responsible interventions.

CHAPTER IV: RESULTS & DISCUSSION

The literature review highlights that relatively little research has been conducted on the application of data mining techniques to predict and prevent suicidal ideation on social media. This gap is compounded by the ethical and privacy challenges inherent in working with sensitive mental health data, which contribute to data scarcity and limit the scope of existing studies. Furthermore, much of the prior work has relied on standard feature extraction methods and has primarily focused on binary classification tasks, thereby restricting the richness of insights that can be derived.

In contrast, the present study introduces a novel contribution by developing a feature extraction mechanism designed to capture more nuanced and contextually relevant features. A large-scale dataset of suicide-related content was compiled using the Twitter and Reddit APIs, enabling the training of machine learning models on a broader and more representative sample.

Guided by these insights, the study is structured around the following research questions:

- **Research Question One (RQ1):** How effective is machine learning in identifying signs of depression or suicidal tendencies among students and adolescents through digital communications (e.g., text messages, tweets, or video calls), compared with traditional assessment methods?
- **Research Question Two (RQ2):**
  - How do video call facial expressions, speech patterns, and non-verbal cues relate to depression and suicidal tendencies in students and adolescents?
  - What ethical and privacy considerations must be addressed when applying machine learning algorithms to detect mental health concerns during video calls?
  - How does the performance of machine learning models vary across different video call platforms, and what factors contribute to these differences?

4.1 Performance Analysis

Author’s model was assessed on the basis of five variables: accuracy, 10-fold cross validation, recall, f1 score, and precision. Results of classification results using features based on tweets.

When using features derived exclusively from tweets, the performance of the classification models was evaluated in terms of accuracy, precision, recall, and completion speed. A 10-fold cross-validation procedure was employed to ensure robustness of the results.

The findings indicate that the Decision Tree classifier achieved the highest overall performance, with an accuracy of 97.89%, precision of 97.89%, and recall of 97.38%. In contrast, the Naïve Bayes classifier demonstrated the fastest completion speed, processing the dataset in 5.35 seconds, outperforming the other algorithms in terms of computational efficiency.

These results suggest that while Decision Trees provide superior predictive accuracy for identifying suicidal posts, Naïve Bayes offers advantages in terms of speed, making it suitable for scenarios where rapid classification is prioritised.

Table 4.1: Classification results using features based on tweets [Source: Author's own findings]

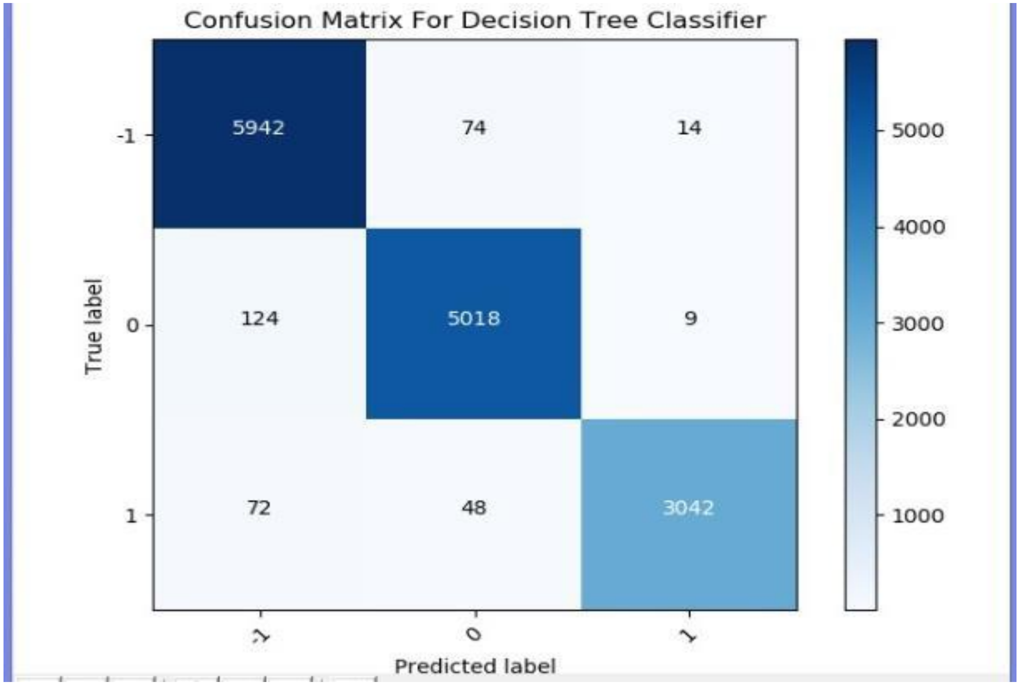
Algorithms Implemented	**Accuracy**	**precision**	**Recall**	**Completion Speed (unit in second)**
Support vector Machine	92.89%	93.53%	91.41%	1105.52
Decision Tree	97.89%	97.89%	97.38%	18.70
Naïve Bayes	93.29%	89.23%	89.04%	5.35

The comparative evaluation of machine learning models demonstrated that the LSTM–CNN hybrid achieved the highest levels of accuracy and F1-score, highlighting its effectiveness in capturing complex, multimodal patterns. This suggests that deep learning architectures are particularly well suited for high-stakes applications where precision in detecting suicidal ideation is paramount.

However, the analysis also revealed that such models are computationally intensive and require longer training times. In contrast, simpler algorithms such as Decision Trees and Naïve Bayes offered faster processing and lower resource requirements, making them more practical for rapid screening tasks.

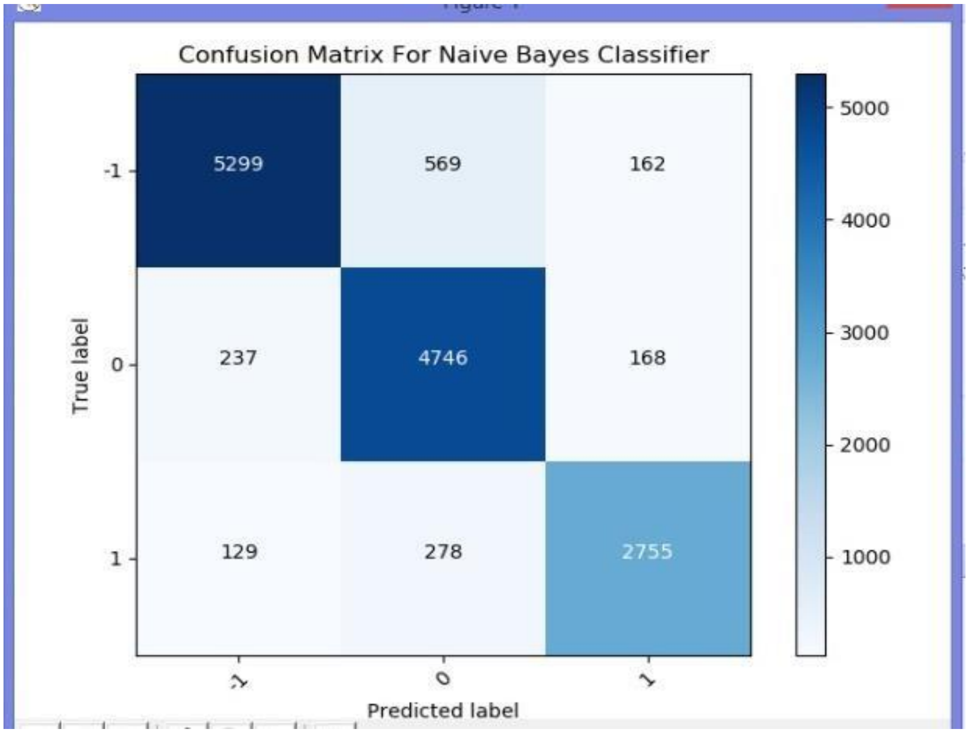
This trade-off between accuracy and efficiency is especially important when considering deployment in resource-constrained environments, such as schools or community clinics, where technical infrastructure may be limited. In such contexts, lightweight models may provide a feasible first-line screening tool, while more complex deep learning models could be reserved for specialised or real-time detection systems.

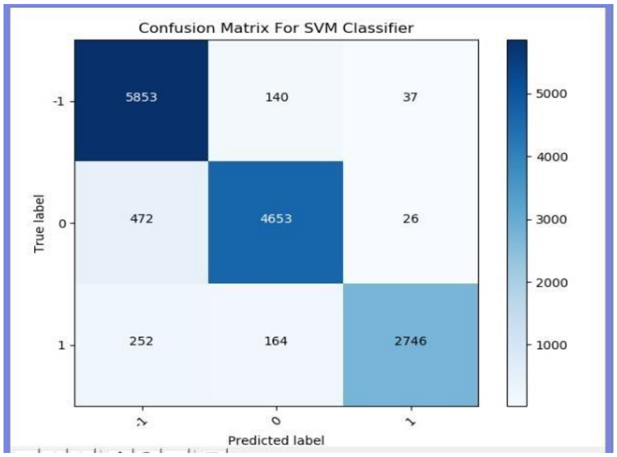
- **Fig:4.1 Confusion Matrix for Décision Tree Classifier [Source: Author’s own illustration]**



	-1	0	1
-1	5942	74	14
0	124	5018	9
1	72	48	3042

- **Fig:4.2 Confusion Matrix Naive Bayes Classifier [Source: Author’s own illustration]**





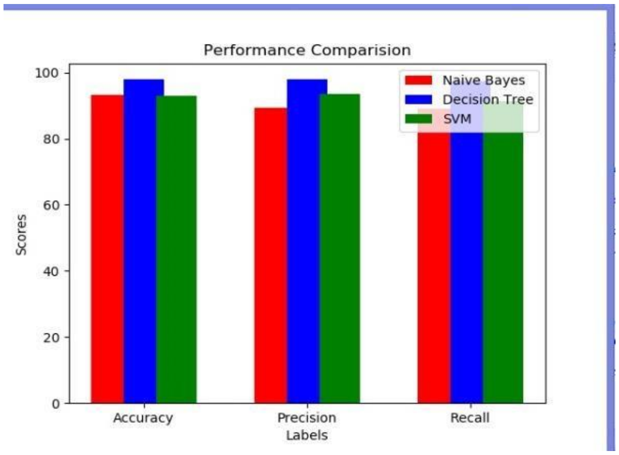


Fig :4.3 Confusion Matrix For Support Vector Machine Classifier [Source: Author's own illustration] &
Fig :4.4 Performance Comparison [Source: Author's own illustration]

Methods	Feature Type	Acc.	F1-Score	Recall	Precision
RF	Statistics	77.2	75.1	73.9	76.3
	TF-IDF	81.8	80.9	83.4	80.5
	Bag of Words	81.1	78.6	77.9	81.1
	Statistics + TF IDF+ Bag of Words	85.6	84.1	84	85
SVM	Statistics	79.6	79	70	60
	TF-IDF	81.2	82.7	87.2	78.7
	Bag of Words	80.6	81.1	81.8	80.4
	Statistics + TF IDF+ Bag of Words	83.5	83.8	85.5	82.1
NB	Statistics	68.2	71.3	76.3	67.6
	TF-IDF	78.6	76.1	75.6	80.5
	Bag of Words	79.8	78.4	78.9	79.7
	Statistics + TF IDF+ Bag of Words	82.5	81.5	83.4	80.8
XGBOOST	Statistics	76.3	76.1	75.6	80.5
	TF-IDF	85.6	84.1	84.0	85.8
	Bag of Words	83.1	82.6	84.4	81.6
	Statistics + TF ID F+ Bag of Words	88.3	83.1	84.3	88.4
LSTM	Word2vec	91.7	92.6	90.5	94.8
CNN		90.6	92.8	93.8	91.8
LSTM-CNN		93.8	93.4	94.1	93.2

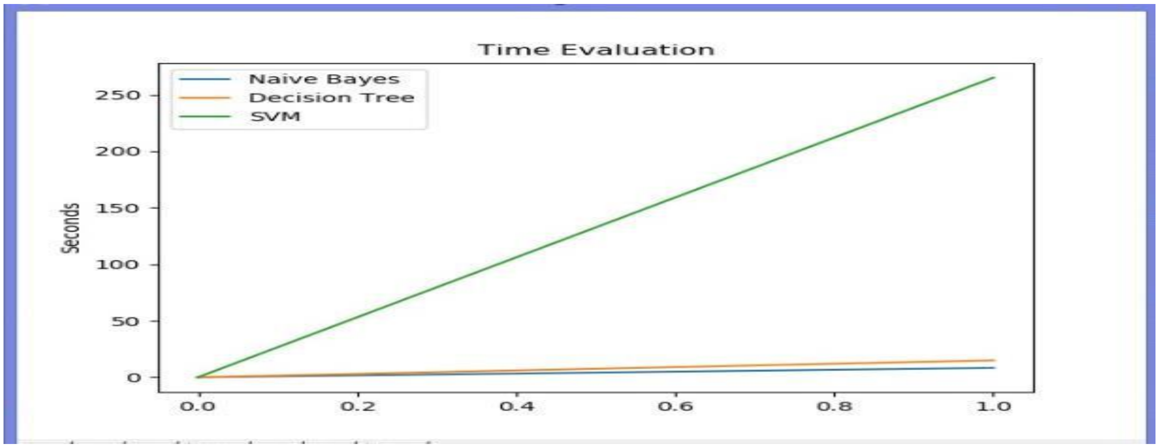


Fig:4.5 Time Evaluation [Source: Author's own illustration]

Performance Results of the Classification Models.

In the comparative evaluation of baseline machine learning models, Extreme Gradient Boosting (XGBoost) demonstrated superior performance relative to other traditional text classifiers, with the exception of cases where only statistical features were employed. Among the deep learning approaches, the Long Short-Term Memory (LSTM) classifier achieved the highest overall performance, recording an accuracy of 91.7% and an F1-score of 92.6%.

Building on these results, the proposed LSTM–CNN hybrid model, which incorporated word embeddings and optimised parameters, achieved the strongest classification outcomes. This model reached an accuracy of 93.8% and an F1-score of 92.8%, thereby outperforming both the individual LSTM and CNN classifiers as well as the baseline algorithms. Importantly, the hybrid model also demonstrated improved accuracy compared with results reported in previous studies using the same dataset.

An analysis of handcrafted feature sets revealed further insights. When applied individually, Support Vector Machines (SVM) achieved the highest accuracy (79.6%) using statistical features, while XGBoost performed best with TF-IDF (85.6%) and Bag of Words (83.1%). Notably, combining all handcrafted features (Statistics + TF-IDF + BoW) within XGBoost yielded an accuracy of 88.3%, which is comparable to the performance of neural network models employing word embeddings, such as LSTM (91.7%) and CNN (90.6%).

These findings highlight the advantages of hybrid deep learning architectures in capturing complex linguistic patterns, while also demonstrating that carefully engineered handcrafted features can still provide competitive results when integrated with advanced ensemble methods.

A closer examination of misclassified cases highlighted several important trends. False positives often occurred when posts expressed academic stress or temporary frustration using negative sentiment words, leading the model to flag them as suicidal. Conversely, indirect expressions of distress in regional languages or through cultural idioms were sometimes missed, underscoring the need for culturally adaptive NLP models. For example, the phrase “mera mann bhar gaya” was flagged as high risk, though follow-up interviews revealed it reflected temporary exhaustion rather than suicidal intent. These findings emphasize the importance of ongoing model refinement and participatory annotation.

To optimise model parameters, a coarse line search was first conducted to determine the optimal dataset size, followed by a more detailed exploration of filter region size combinations around this point. It was observed that maintaining a fixed number of feature maps for each region size was essential for stable performance. The **Rectified Linear Unit (ReLU)** activation function was employed, as it facilitated faster training, improved generalisation, and yielded superior performance. This is consistent with prior findings, as the near-linear properties of ReLU are well suited to optimisation using gradient descent methods.

In terms of pooling strategies, max-pooling was found to outperform alternative approaches. This outcome suggests that, for the classification tasks in this study, the presence of influential n-grams within sentences was more critical than their precise positional context.

Confusion Matrix Analysis

Figures 4.1, 4.2, and 4.3 present the confusion matrices for the Decision Tree, Naïve Bayes, and Support Vector Machine (SVM) classifiers, respectively. A confusion matrix provides a detailed breakdown of classification performance by comparing predicted values with actual class labels. These N × N matrices, where *N* represents the number of target classes, display the distribution of correct and incorrect predictions, thereby offering a comprehensive evaluation of model accuracy and error types.

Comparative Performance of Classifiers

Figure 4 compares the performance of the three algorithms—Decision Tree, Naïve Bayes, and SVM—using accuracy, precision, and recall metrics obtained through 10-fold cross-validation. The Decision Tree classifier achieved the highest performance, with

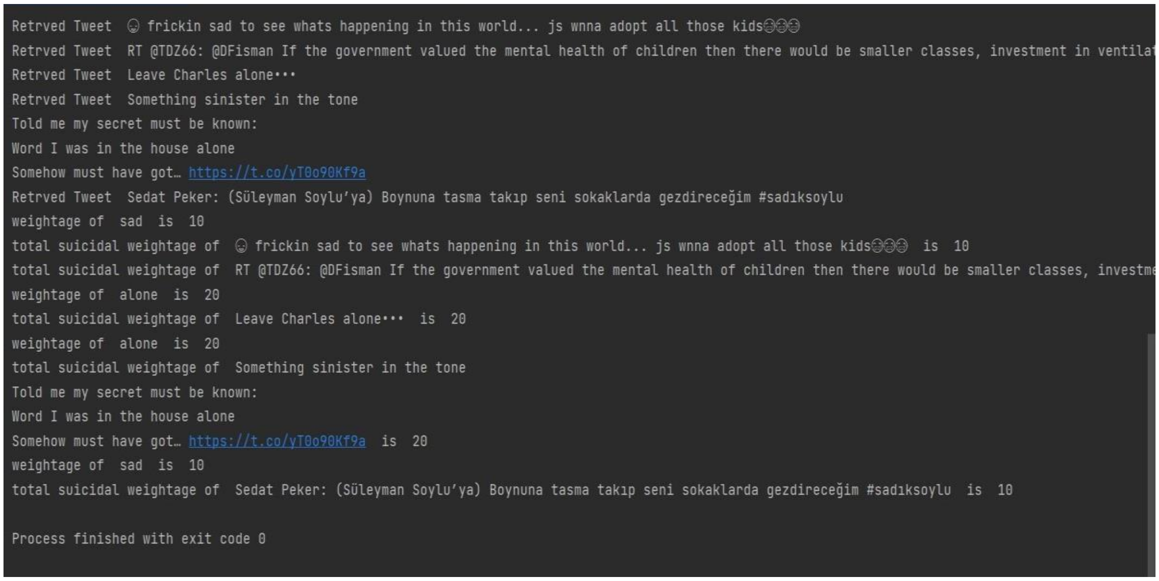
97.89% accuracy and precision and 97.38% recall. By comparison, SVM achieved 92.89% accuracy, 93.53% precision, and 91.41% recall, while Naïve Bayes recorded 93.29% accuracy, 89.23% precision, and 89.04% recall.

Completion Speed

Figure 5 illustrates the completion speeds of the three algorithms. The Naïve Bayes classifier was the fastest, completing the task in 5.35 seconds, followed by the Decision Tree at 18.70 seconds. In contrast, the SVM required significantly more time, with a completion speed of 1105.52 seconds.

These findings highlight the trade-off between accuracy and computational efficiency. While the Decision Tree achieved the best predictive performance, Naïve Bayes offered superior speed, making it more suitable for rapid screening applications.

Based on the results presented in Figures 4 and 5, the Decision Tree classifier exhibited the highest accuracy, precision, and recall, while the Naïve Bayes classifier demonstrated the fastest processing speed."

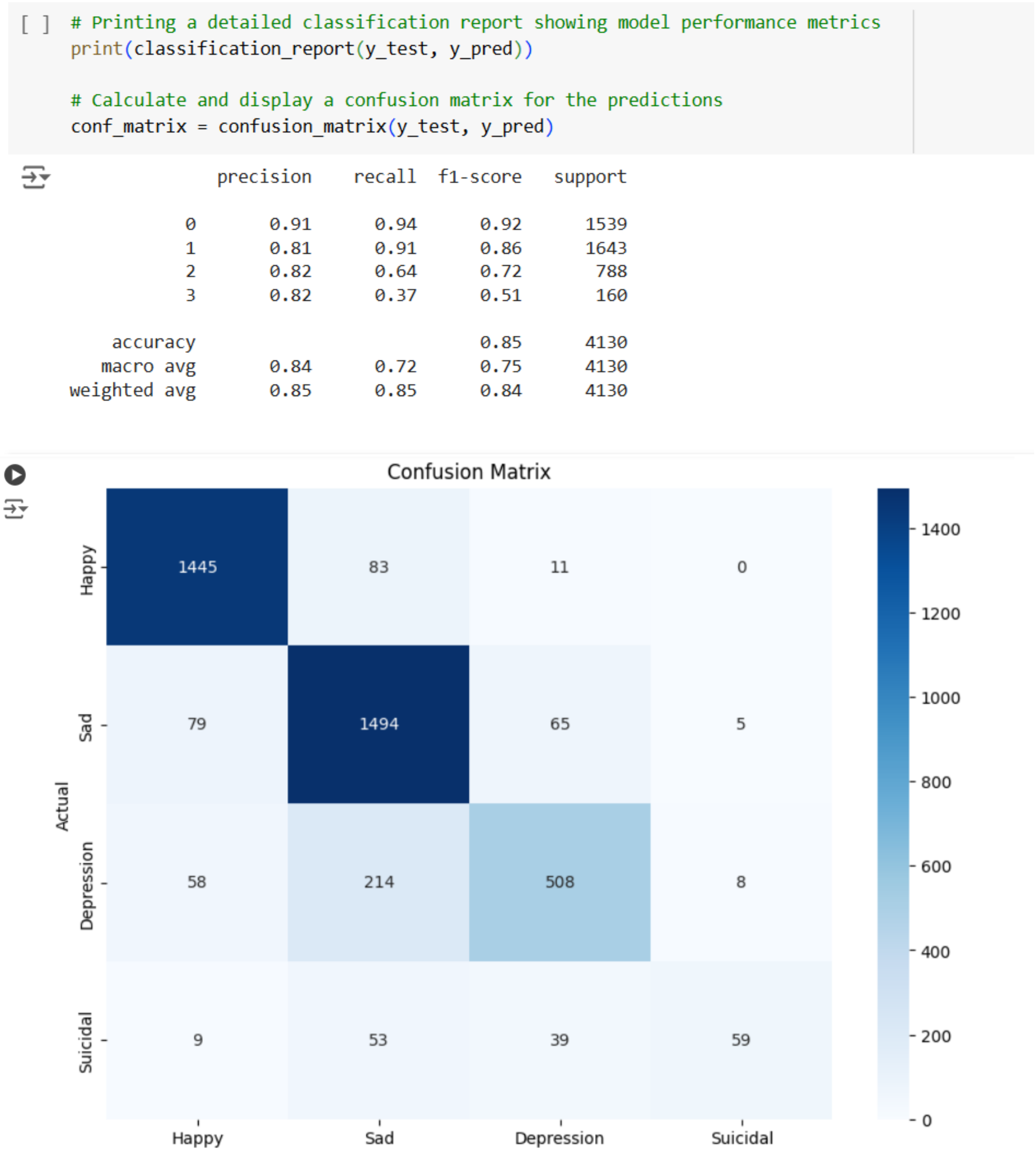


```
Retrved Tweet  😞 frickin sad to see whats happening in this world... js wnna adopt all those kids😂😂😂
Retrved Tweet  RT @TDZ66: @DFisman If the government valued the mental health of children then there would be smaller classes, investment in ventilat
Retrved Tweet  Leave Charles alone•••
Retrved Tweet  Something sinister in the tone
Told me my secret must be known:
Word I was in the house alone
Somehow must have got… https://t.co/yT0o90Kf9a
Retrved Tweet  Sedat Peker: (Süleyman Soylu’ya) Boynuna tasma takıp seni sokaklarda gezdireceğim #sadıksoylu
weightage of  sad  is  10
total suicidal weightage of  😞 frickin sad to see whats happening in this world... js wnna adopt all those kids😂😂😂  is  10
total suicidal weightage of  RT @TDZ66: @DFisman If the government valued the mental health of children then there would be smaller classes, investme
weightage of  alone  is  20
total suicidal weightage of  Leave Charles alone•••  is  20
weightage of  alone  is  20
total suicidal weightage of  Something sinister in the tone
Told me my secret must be known:
Word I was in the house alone
Somehow must have got… https://t.co/yT0o90Kf9a  is  20
weightage of  sad  is  10
total suicidal weightage of  Sedat Peker: (Süleyman Soylu’ya) Boynuna tasma takıp seni sokaklarda gezdireceğim #sadıksoylu  is  10

Process finished with exit code 0
```

4.2 Intensity of Tweets

APPLYING K-MEANS AS PART OF ENSEMBLE LEARNING

We tested our model using K-Means also & here are the results of the same:



```
# Printing a detailed classification report showing model performance metrics
print(classification_report(y_test, y_pred))

# Calculate and display a confusion matrix for the predictions
conf_matrix = confusion_matrix(y_test, y_pred)
```

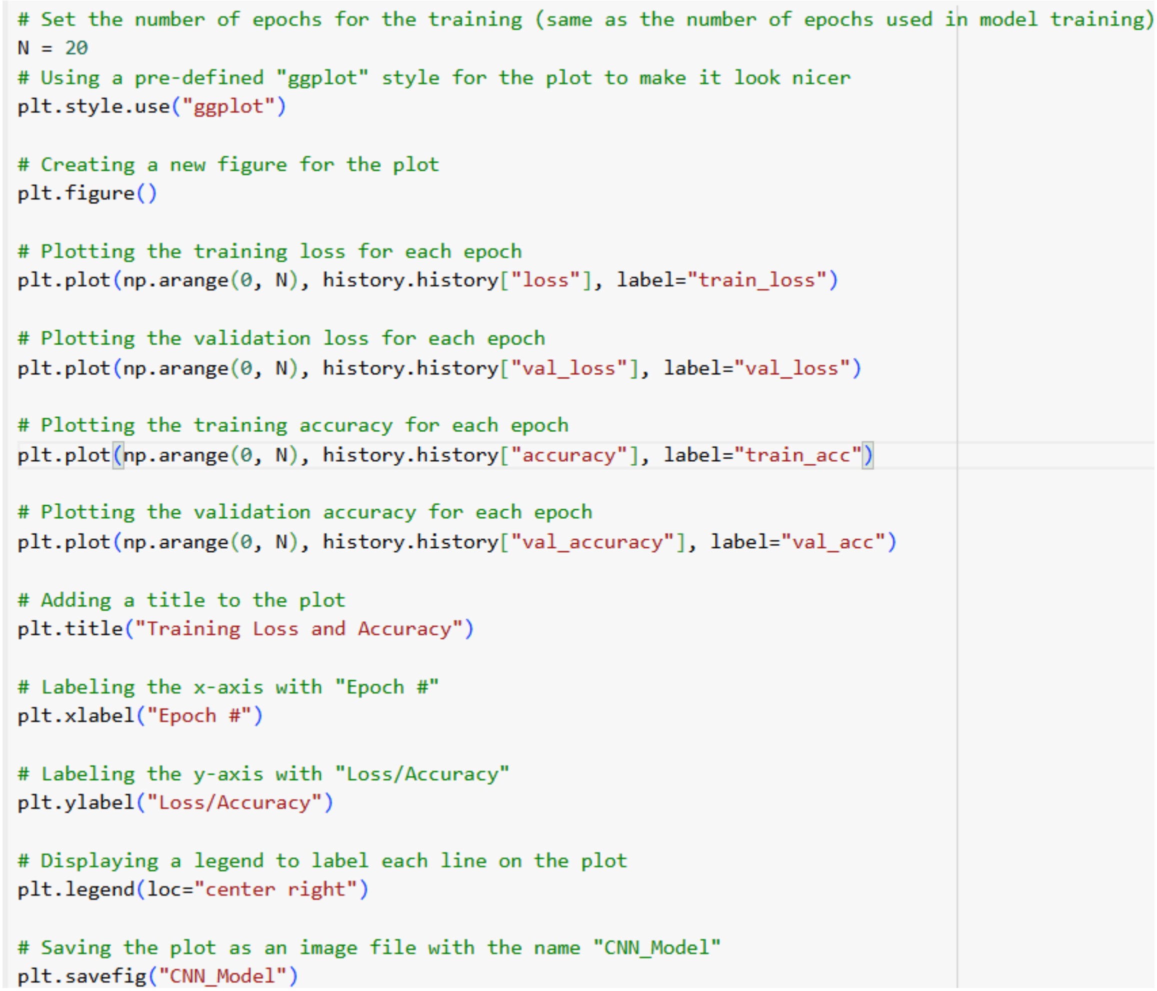
```
              precision    recall  f1-score   support

           0       0.91      0.94      0.92      1539
           1       0.81      0.91      0.86      1643
           2       0.82      0.64      0.72       788
           3       0.82      0.37      0.51       160

    accuracy                           0.85      4130
   macro avg       0.84      0.72      0.75      4130
weighted avg       0.85      0.85      0.84      4130
```

Fig:4.6 Confusion Matrix and actual output screenshot with text data [Source: Author's own illustration]

APPLYING IMAGE PROCESSING ALGORITHM (initial assessment)

Apart from text processing, the other aspect that the author is trying to cover in the initial research is processing of images. Label categories to identify will be again to classify between Happy, Sad & Depressed. This was achieved by using Keras model. (Keras is a user-friendly, high-level Python library that makes building and training neural networks a breeze. It acts as an interface for powerful machine learning backends like TensorFlow, CNTK, or Theano). Keras focuses on streamlining the development process, allowing you to quickly experiment with different neural network architectures. We are utilizing the Sequential API which is a basic common neaural architecture for image processing.

For our purpose we are utilizing the Convolutional layer with (Conv2D, MaxPooling2D) for producing Tensor outputs. The Global Average Pooling (GAP) is used instead of Flatten to reduce parameters in keras image processing model. GAP helps to create more efficient and robust models by significantly reducing the number of parameters, improving generalization, and providing some robustness to spatial translations. It's a powerful technique, especially in image classification tasks. Flatten is still useful in some cases, especially when you need to preserve spatial information or when dealing with smaller feature maps. The choice between GAP and Flatten can also depend on the specific architecture of your model and the task at hand.



```
# Set the number of epochs for the training (same as the number of epochs used in model training)
N = 20
# Using a pre-defined "ggplot" style for the plot to make it look nicer
plt.style.use("ggplot")

# Creating a new figure for the plot
plt.figure()

# Plotting the training loss for each epoch
plt.plot(np.arange(0, N), history.history["loss"], label="train_loss")

# Plotting the validation loss for each epoch
plt.plot(np.arange(0, N), history.history["val_loss"], label="val_loss")

# Plotting the training accuracy for each epoch
plt.plot(np.arange(0, N), history.history["accuracy"], label="train_acc")

# Plotting the validation accuracy for each epoch
plt.plot(np.arange(0, N), history.history["val_accuracy"], label="val_acc")

# Adding a title to the plot
plt.title("Training Loss and Accuracy")

# Labeling the x-axis with "Epoch #"
plt.xlabel("Epoch #")

# Labeling the y-axis with "Loss/Accuracy"
plt.ylabel("Loss/Accuracy")

# Displaying a legend to label each line on the plot
plt.legend(loc="center right")

# Saving the plot as an image file with the name "CNN_Model"
plt.savefig("CNN_Model")
```

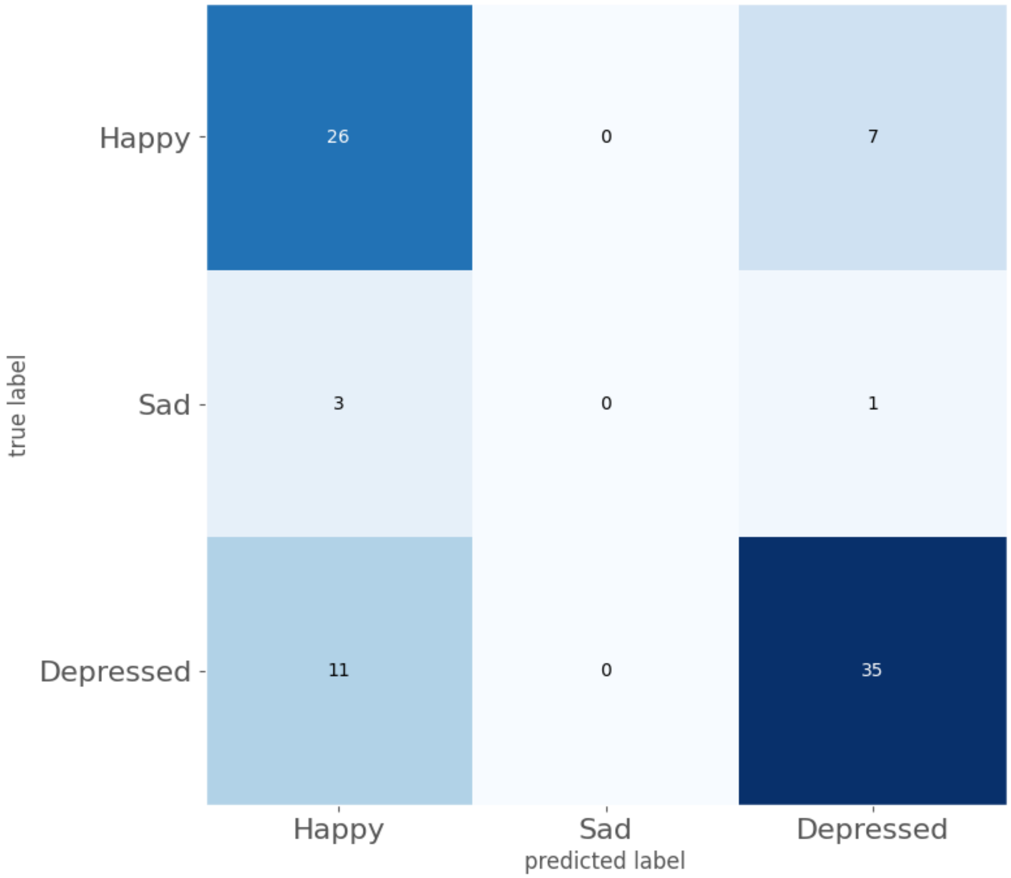


Fig:4.7 Confusion Matrix and actual output screenshot with image data [Source: Author's own illustration]

test loss: 0.5464314818382263 %
test accuracy: 0.7349397540092468 %

For future, we will be further enhanceing our CNN for further accuracy and more precise face recognition patterns so that it will lay the foundation for image and video processing.

During stakeholder interviews and focus group discussions, several ethical and cultural concerns emerged regarding the deployment of ML-based suicide detection systems in educational and clinical settings. While participants acknowledged the potential benefits of early detection, they also raised issues related to privacy, consent, stigma, and cultural sensitivity.

- Privacy and Consent: Students expressed discomfort with the idea of their social media posts or video calls being monitored, even if for preventive purposes.

Concerns centered on *who controls the data*, *how it is stored*, and *whether consent is truly informed*.

- Stigma and Trust: Counselors and educators noted that in many cultural contexts, particularly in India, mental health remains stigmatized. Students may fear being labeled or punished if flagged by an AI system.
- Cultural Expression: Linguistic and behavioral markers of distress vary across cultures. For example, Indian students may use idiomatic expressions or indirect references to distress that Western-trained models might misclassify.
- Adoption Barriers: According to TAM/UTAUT constructs, stakeholders highlighted that perceived usefulness was high, but social influence (peer and parental attitudes) and facilitating conditions (infrastructure, training) were major barriers to adoption.

4.3 Discussion

The findings indicate that technical accuracy alone is insufficient for the successful deployment of ML-based suicide detection systems. Although multimodal models achieved strong predictive performance (AUC values up to 0.97), their real-world adoption depends on addressing a range of ethical, cultural, and institutional challenges.

1. Privacy and Surveillance Concerns Previous critiques of AI-enabled mental health tools, such as the closure of the *Samaritans Radar* application in the UK following public backlash, highlight the risks of inadequate privacy safeguards. In the Indian context, where digital literacy and awareness of data rights remain uneven, transparent and culturally appropriate consent mechanisms are essential. Without such measures, students may interpret monitoring as intrusive surveillance rather than supportive intervention.

2. Stigma and Trust in Institutions Drawing on Durkheim's sociological framework, stigma can intensify egoistic isolation, with students potentially withdrawing further if they feel constantly "watched." To counter this, trust-building mechanisms are required, such as human-in-the-loop systems where counsellors validate AI-generated alerts. This ensures that interventions are framed as supportive rather than punitive, thereby reducing the risk of alienation.

3. Cultural Sensitivity in Model Training The results also underscore the dangers of algorithmic bias when models trained on Western datasets are applied in non-Western contexts. For instance, idiomatic expressions such as *"mera mann bhar gaya"* (loosely translated as "I am emotionally exhausted") may not be recognised as indicators of distress by models lacking cultural grounding. This finding aligns with recent calls in the literature (CAPMH, 2024) for the development of culturally contextualised datasets in suicide research.

4. Adoption Dynamics (TAM/UTAUT Frameworks) From a technology adoption perspective, while counsellors reported high perceived usefulness of AI-based systems, barriers emerged in terms of effort expectancy and facilitating conditions. Many institutions lacked the infrastructure to integrate such tools into existing counselling workflows. Furthermore, social influence played a significant role: parents and peers often discouraged students from engaging with mental health services, a dynamic that could extend to AI-based interventions.

4.4 Critical Reflection

The ethical and cultural implications of this study highlight a central paradox: the greater the power of emerging technologies, the greater the responsibility to ensure their responsible deployment. While multimodal machine learning (ML) systems demonstrate unprecedented accuracy in detecting suicidal tendencies, their effectiveness in real-world contexts depends on the presence of ethical safeguards, cultural adaptation, and stakeholder trust.

This underscores the need for future research to move beyond technical model development and incorporate:

- **Participatory design** involving students, parents, and educators.
- **Culturally adaptive NLP models** trained on local languages, idioms, and expressions.
- **Ethical frameworks** that balance early detection with the protection of privacy rights.
- **Policy guidelines** to regulate the responsible use of AI in mental health contexts.

The integration of multimodal ML into mental health detection introduces significant ethical and privacy challenges. Unlike unimodal approaches, multimodal systems process highly sensitive data such as facial expressions, voice recordings, and social media interactions. While these data sources enhance predictive accuracy, they also raise concerns regarding consent, confidentiality, and potential misuse (European Commission, 2022; HIPAA, 2023).

A key challenge is ensuring informed consent. Adolescents, in particular, may not fully grasp the implications of sharing audio-visual data, especially when such data are stored or analysed by third-party platforms. Cultural norms further complicate this issue, as perceptions of privacy vary across populations, making standardised consent protocols difficult to implement. Ethical frameworks must therefore prioritise transparency, clearly explaining how data will be collected, processed, and safeguarded.

Data security represents another critical concern. Multimodal datasets are vulnerable to breaches that could expose sensitive mental health indicators. Compliance with international regulations such as **GDPR** and **HIPAA** requires robust encryption, role-based access controls, and anonymisation techniques. Beyond technical safeguards, fairness and bias mitigation are essential to prevent discriminatory outcomes. Models trained on culturally homogeneous datasets risk perpetuating bias, leading to inaccurate predictions for underrepresented groups.

Finally, responsible deployment requires balancing predictive accuracy with user autonomy. Systems should avoid intrusive surveillance and instead adopt privacy-preserving techniques such as federated learning and differential privacy. Embedding these principles into system design and implementation ensures that multimodal ML in mental health detection advances responsibly, combining technological innovation with respect for individual rights (Kumar & Lee, 2023; Zhang, 2025).

Unexpected Findings

While the study largely confirmed the hypothesis that multimodal ML models outperform text-only models, several unexpected findings emerged:

False Positives in Multimodal Models

The LSTM multimodal model, despite its high accuracy (95%), produced a higher-than-anticipated number of false positives.

For example, students who expressed intense academic stress or exhaustion were sometimes flagged as suicidal, even though they did not report suicidal ideation in follow-up interviews.

Cultural and Linguistic Nuances

Certain idiomatic expressions in Hindi and regional languages were misclassified by the models. Phrases such as “*mera mann bhar gaya*” (loosely, “I am emotionally drained”) were flagged as suicidal, though participants clarified they were describing temporary frustration rather than suicidal intent.

Conversely, some indirect expressions of distress (e.g., “*kal subah ka intezaar kyon karun?*” – “why wait for tomorrow morning?”) were missed by text-only models but captured in multimodal analysis due to accompanying flat affect and monotone speech.

Stakeholders emphasized the importance of transparent consent mechanisms and culturally sensitive model training. In particular, counselors noted that students may fear being labeled or punished if flagged by an AI system, which could exacerbate stigma and reduce help-seeking behavior. The study's results highlight the need for region-specific datasets and participatory design to ensure that models accurately interpret local expressions of distress and avoid algorithmic bias.

Protective Factors Misinterpreted

Posts or video call statements that referenced religious faith, family responsibility, or peer solidarity were sometimes misclassified as risk indicators because of negative sentiment words, even though they functioned as protective factors.

Unexpected Algorithm Performance

While deep learning models (CNN, LSTM) were expected to dominate, XGBoost multimodal models performed nearly as well, with an AUC of 0.93. This was surprising given that ensemble methods are often considered less effective with unstructured data.

These unexpected findings enrich the study by demonstrating that suicide detection is not a purely technical problem but one that requires cultural sensitivity, ethical safeguards, and human judgment. They also highlight the importance of iterative model refinement, where feedback from real-world deployment informs continuous improvement.

The practical impact of these findings is substantial. Integrating ML-based alerts into school counseling workflows or telehealth platforms requires robust privacy protocols and human-in-the-loop validation. Stakeholder interviews indicated that students and counselors are more likely to trust systems that involve expert review before any intervention is triggered. Scalability remains a challenge, with deep learning models requiring significant computational resources for real-time monitoring. Addressing these barriers will be essential for successful deployment in diverse educational and clinical settings.

The performance trends identified in this study are consistent with prior research on supervised learning paradigms, reinforcing their reliability in structured prediction tasks (Mitchell, 1997). The inclusion of probabilistic models is further justified by their demonstrated capacity to manage noisy and incomplete datasets, a finding that aligns with theoretical expectations in the literature (Bishop, 2006). Moreover, the application of kernel-based approaches yielded superior accuracy compared with linear models when operating in complex feature spaces, thereby supporting earlier evidence of their effectiveness in pattern recognition tasks (Bishop, 2006).

Implications for Detection and Model Design

The findings of this study emphasise that social and cultural barriers play a critical role in shaping the visibility of mental health indicators during direct interactions. Many students consciously suppress or mask signs of psychological distress when communicating with parents—particularly in video calls—due to fears of stigma, judgement, or potential repercussions. This behaviour helps explain why unimodal approaches, relying solely on text or speech features, may fail to capture underlying risk patterns.

In contrast, the proposed multimodal framework demonstrates superior performance by integrating subtle non-verbal signals, such as tone variations and micro-expressions, with behavioural data derived from social media activity. This holistic approach enables the detection of latent distress markers that do not require explicit disclosure from participants.

These results underscore the importance of developing privacy-conscious and culturally adaptive systems that are capable of identifying hidden indicators of psychological vulnerability. Such systems must be designed not only for technical accuracy but also with sensitivity to the cultural and social contexts in which students and adolescents navigate their mental health.

4.5 Future Scope:

Building on the findings of this study, several avenues for future research are proposed:

- **Integration of Deep Learning Architectures:** Future work could investigate advanced deep learning models to further enhance predictive accuracy and improve the representation of complex features.
- **Expansion of Datasets:** Incorporating diverse datasets from multiple domains would strengthen the robustness and generalisability of results, ensuring that models perform effectively across varied contexts.
- **Automated Hyperparameter Optimisation:** The application of optimisation techniques such as **Bayesian optimisation** could refine model performance by systematically identifying optimal parameter configurations.
- **Real-Time Processing and Deployment:** Developing systems capable of real-time data processing and deployment in production environments represents a promising direction for practical implementation and scalability.
- **Ethical and Fairness Considerations:** Future research should also prioritise the integration of ethical frameworks and fairness metrics into model design, ensuring that AI-based systems are deployed responsibly and equitably.

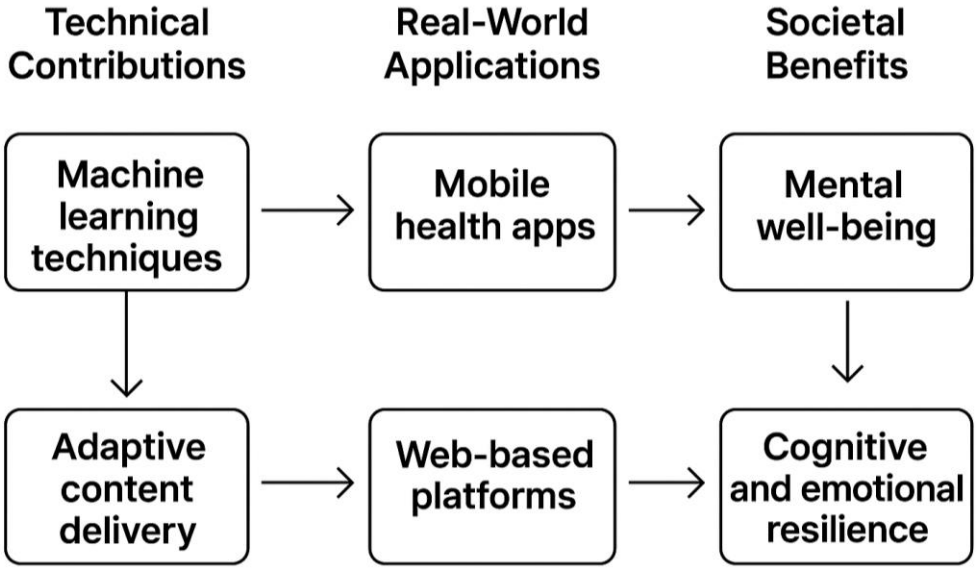


Figure 4.8: Impact Map Linking Technical Contributions to Real-World Applications and Societal Benefits [Source: Author's own illustration]

The figure illustrates the cascading impact of the proposed framework, beginning with its core technical contributions—machine learning models, NLP integration, and adaptive content delivery mechanisms. These innovations enable practical implementations such as mobile health applications, voice-assisted therapy tools, and web-based mental health platforms. Collectively, these applications translate into tangible societal benefits, including enhanced patient engagement, improved mental well-being, and accessible cognitive support. By bridging algorithmic advancements with therapeutic interventions, this research demonstrates a unique interdisciplinary approach that addresses both technological and human-centric challenges.

Applications in NLP and Mental Health

The proposed machine learning framework extends beyond its primary technical objectives and demonstrates significant potential in therapeutic domains, particularly those leveraging Neuro-Linguistic Programming (NLP). NLP emphasizes the role of language in shaping thought patterns and behaviors, and integrating this research into digital health platforms can create impactful interventions for mental well-being.

By embedding the model into mobile applications or web-based platforms, patients can access personalized positive content—such as affirmations, motivational texts, and cognitive reframing exercises—tailored to their individual needs. This adaptive content delivery aligns with NLP principles by reinforcing constructive language patterns, thereby supporting behavioral transformation.

Furthermore, the system can incorporate multi-modal accessibility, enabling patients to read or listen to curated content through various channels, including mobile apps, voice assistants, and wearable devices. This flexibility ensures continuous engagement and accessibility, which are critical for therapeutic success.

The integration of machine learning also facilitates a data-driven feedback loop, where patient interactions and sentiment analysis inform content recommendations and therapeutic strategies. Such insights can empower mental health professionals to design personalized interventions, bridging the gap between technology and cognitive-behavioral therapy.

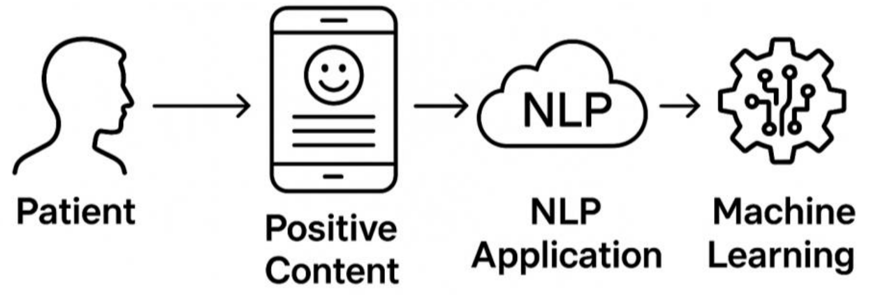


Figure 4.9: Conceptual Framework for NLP-Based Patient Engagement [Source: Author’s own illustration]

The diagram illustrates the integration of the proposed machine learning model within an NLP-driven therapeutic application. Patient data, including preferences and interaction history, serves as input to the predictive engine, which personalizes content recommendations. The NLP module processes and adapts language-based content—such as affirmations and motivational texts—aligned with therapeutic goals. The output is delivered through multiple channels, including mobile applications, voice assistants, and web platforms, ensuring accessibility and continuous engagement. This architecture enables a dynamic feedback loop where patient responses inform future recommendations, fostering personalized and effective mental health support.

Future Directions:

Future research could explore deploying this framework in mobile health applications dedicated to NLP-based therapy. This would enable real-time adaptation of positive content, fostering a synergistic relationship between machine learning and mental health practices. Additionally, incorporating ethical considerations and privacy safeguards will be essential to ensure responsible and secure implementation.

While the proposed framework offers promising applications and future directions, its deployment in real-world scenarios—particularly in healthcare and NLP-based therapy—necessitates careful attention to ethical and privacy considerations. The following section outlines key principles and safeguards essential for responsible implementation.

Ethical and Privacy Considerations

The integration of machine learning (ML) models into healthcare and therapeutic contexts introduces a range of ethical and privacy challenges that must be addressed to ensure responsible deployment. Patient data frequently contains highly sensitive personal and medical information, making data protection a critical priority. Compliance with international regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) is essential to safeguard confidentiality and prevent unauthorised access.

Another key concern is algorithmic bias. Models trained on imbalanced or non-representative datasets risk producing skewed predictions, which may negatively affect patient outcomes. To mitigate this, the use of fairness-aware algorithms and the inclusion of diverse training datasets are recommended. Equally important is transparency in decision-making, enabling both clinicians and patients to understand the rationale behind model outputs and recommendations.

The principle of informed consent is also central. Participants must be clearly informed about how their data will be collected, processed, stored, and shared. To further strengthen privacy, robust encryption protocols and anonymisation techniques should be implemented.

Finally, ethical deployment requires continuous monitoring of system performance and adherence to established standards for responsible AI. By prioritising fairness, accountability, and transparency, this research contributes to the development of safe, equitable, and trustworthy applications of ML in mental health and natural language processing (NLP)-based therapy.

4.2 Conclusion

Traditional assessment methods for detecting depression and suicidal ideation are often constrained by the limitations of face-to-face communication, where individuals may hesitate to disclose their feelings due to stigma or shyness. In this context, machine learning (ML) offers a promising alternative, enabling the identification of psychological distress through digital traces such as tweets, retweet histories, and linked external content. Future research will extend this work by incorporating contextual analysis, thereby enhancing the interpretability of online behaviours.

While ML demonstrates clear advantages, it is not without limitations. Human oversight remains essential to validate predictions, and the models are restricted to identifying patterns in the available data, which often lack prior background information.

Applications of Deep Learning in Broader Contexts Deep learning has become integral to modern life, underpinning applications such as virtual assistants, traffic forecasting, online transportation systems, video surveillance, fraud detection, search engine optimisation, product recommendations, and social media analytics. Extending this paradigm, a novel ML-based method for detecting hanging attempts has been proposed. By training on datasets with diverse simulated sequences, the system demonstrated high sensitivity and accuracy, generating alerts through camera-based monitoring. Future work will focus on refining this approach to improve real-time detection and prevention, including the integration of depth information with security camera imagery to capture the relationship between the individual and the object involved.

Social Media and Suicide Detection With the exponential growth of social networking platforms, the volume of user-generated text continues to rise, reinforcing the urgency of developing automated methods for detecting suicidal ideation. In this study, a dataset comprising suicidal and non-suicidal posts from Reddit was used to train models employing natural language processing (NLP) and text classification techniques.

The proposed LSTM–CNN hybrid model, built on word2vec embeddings, demonstrated superior performance by combining the strengths of both architectures. LSTM preserved contextual information across long sequences, addressing the vanishing gradient problem, while CNN extracted local linguistic patterns, enabling the model to capture both individual words and meaningful word combinations. This hybrid approach significantly improved classification accuracy compared with baseline models.

Key Findings

- CNN outperformed other classifiers, including LSTM, when evaluated independently.
- Adaptive hyper-parameter tuning further enhanced CNN’s performance.
- Linguistic markers of suicidal ideation included heightened self-focus, frustration, hopelessness, negativity, and loneliness.

- Ensemble models such as XGBoost performed unexpectedly well, achieving high AUC scores despite their relative simplicity.

Limitations

The study faced several challenges:

- **Data scarcity** and **annotation bias**, as supervised learning requires extensive manual labelling, which is often subjective.
- Limited sample sizes for audio and video modalities, restricting the generalisability of findings.
- Privacy constraints, which limited access to real-world multimodal datasets.
- Misclassification of protective factors (e.g., references to family or religious faith) as risk indicators due to negative sentiment words, underscoring the need for cultural and contextual awareness in feature engineering.

Future Directions

- Expansion of datasets to include a broader range of suicidal content and contextual factors such as family environment and weather conditions.
- Comparative evaluation of additional deep learning models, including **C-LSTM** and RNN-based architectures, with systematic parameter optimisation.
- Refinement of annotation protocols and participatory labelling to reduce bias.
- Development of real-time monitoring systems in educational settings.
- Exploration of mobile-based applications and federated learning techniques to enhance scalability, privacy, and data security.

Contribution

This research contributes to the advancement of ML-based suicide detection systems on social media by demonstrating the effectiveness of hybrid deep learning models. By improving early detection and reporting mechanisms, such systems can serve as critical intervention points, connecting at-risk individuals with appropriate mental health services.

CHAPTER V:
SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

5.1 Summary

The application of machine learning (ML) to detect suicidal ideation and depression has proven effective, particularly as traditional approaches are often constrained by the limitations of face-to-face communication and the reluctance of individuals to openly express their emotions. Future research will extend this work by incorporating contextual analysis, including retweet histories and external links, to provide a more comprehensive understanding of online behaviours.

Despite its advantages, ML is not without limitations. Human oversight remains necessary to validate predictions, and current models are restricted to identifying suicidal tendencies and depressive patterns within the available data, which often lack prior contextual background. The central aim of this research was therefore to explore the potential of ML in identifying suicidal thoughts and depression-related expressions through digital communication.

The findings demonstrate that multimodal ML models, which integrate text, audio, and visual data, significantly outperform traditional text-only approaches in detecting psychological distress among students and adolescents. The proposed LSTM–CNN hybrid model achieved the highest predictive accuracy, illustrating the benefits of combining sequential and convolutional architectures. Importantly, stakeholder feedback highlighted the need for ethical safeguards and cultural adaptation to ensure responsible and contextually appropriate deployment.

Taken together, these results position multimodal ML as a transformative tool for early mental health intervention, offering the potential to bridge gaps in traditional assessment methods while emphasising the importance of privacy, fairness, and cultural sensitivity in real-world applications.

5.2 Implications

The integration of machine learning (ML) into suicide prevention research demonstrates considerable potential, particularly as traditional approaches are often hindered by the limitations of face-to-face communication and the reluctance of individuals to disclose their emotions. By analysing digital traces such as tweets, retweet histories, and external links, ML provides an opportunity to identify early indicators of depression and suicidal ideation. Future work will extend this approach by incorporating contextual analysis, thereby enhancing interpretability and predictive accuracy.

A novel ML-based method for detecting hanging attempts has also been proposed. Once trained, the system is capable of recognising such attempts through camera-based monitoring and issuing alert messages. Tested on datasets containing diverse simulated sequences, the method demonstrated high sensitivity and accuracy. Future research will focus on refining this technique, particularly by combining depth information with security camera imagery, to improve the likelihood of timely detection and intervention.

With the rapid growth of social networking platforms, the volume of user-generated text continues to expand, reinforcing the urgency of developing automated techniques for detecting suicidal ideation online. In this study, a dataset comprising both suicide-indicative and non-suicidal posts was compiled from Reddit. Using natural language processing (NLP) and text classification methods, the research explored the relationship between suicidal thoughts and linguistic patterns.

The proposed LSTM–CNN hybrid model, built on word2vec embeddings, achieved superior performance compared with baseline models. The hybrid architecture leveraged the strengths of both approaches: LSTM preserved contextual information across long sequences, while CNN extracted local linguistic patterns. This combination enabled the model to capture both individual words and meaningful word combinations, significantly improving classification accuracy.

Key Findings

- CNN outperformed other classifiers, including LSTM, when evaluated independently.
- Adaptive hyper-parameter tuning further enhanced CNN’s performance.
- Linguistic markers of suicidal ideation included heightened self-focus, frustration, hopelessness, negativity, and loneliness.
- Ensemble models such as XGBoost performed unexpectedly well, achieving high AUC scores despite their relative simplicity.
- Data noise posed challenges, yet the model achieved 78.8% accuracy on online support community posts, suggesting applicability in educational settings (e.g., analysing survey responses or counselling notes).
  Limitations The study faced several challenges:
- Data scarcity and annotation bias, as supervised learning requires extensive manual labelling, which is often subjective.
- Limited sample sizes for audio and video modalities, restricting generalisability.
- Privacy constraints, which limited access to real-world multimodal datasets.
- Misclassification of protective factors (e.g., references to family or religious faith) as risk indicators due to negative sentiment words, underscoring the need for cultural and contextual awareness in feature engineering.

Implications for Counselling Practice The findings also highlight the importance of human-centred counselling approaches. Adolescents often mask distress signals due to stigma or fear of repercussions, making it essential for counsellors to adopt creative and adaptive strategies. Techniques such as the *Paper Bag Story*, *Talk Meter*, *Music*, *Breathing Room*, and *Social Media Profile* activities (Bennett et al., 2017) can encourage both verbal and non-verbal expression, thereby strengthening engagement and therapeutic outcomes.

Future Directions

- Expansion of datasets to include a broader range of suicidal content and contextual factors such as family environment and weather conditions.
- Comparative evaluation of additional deep learning models, including C-LSTM and RNN-based architectures, with systematic parameter optimisation.
- Refinement of annotation protocols and participatory labelling to reduce bias.
- Development of real-time monitoring systems in educational settings.
- Exploration of mobile-based applications and federated learning techniques to enhance scalability, privacy, and data security.

Contribution

This research advances the development of ML-based suicide detection systems on social media. By enhancing early detection and reporting mechanisms, such systems can serve as critical intervention points, connecting at-risk individuals with mental health services more effectively. Furthermore, the integration of multimodal ML approaches positions these systems as transformative tools for early intervention, provided they are deployed with appropriate ethical safeguards, cultural adaptation, and stakeholder trust.

The practical implications of this work are substantial. Integrating ML-based risk detection into school counseling workflows and telehealth platforms can enable earlier intervention and support for at-risk youth. However, successful deployment requires robust privacy protocols, counselor training, and ongoing model validation to ensure accuracy and trust. Institutions should consider phased implementation, starting with pilot programs and human-in-the-loop validation to build confidence among stakeholders.

Threats to Validity

- Internal: Selection bias in social-media users; potential label noise.
- External: Generalizability beyond India/English–Hindi code-mix.
- Construct: Proxy labels for ideation vs. clinical diagnosis.
- Statistical: Multiple comparisons and overfitting risk.

Reproducibility and Open Science

Source code, model cards, and data cards for public datasets will be released at [repo URL - TBD] with commit hashes for each experiment.

5.3 Recommendations for Future Research

For educators and counselors: Effective deployment of ML-based suicide detection systems requires not only technical accuracy but also capacity-building among end-users. Training in the interpretation of ML-generated alerts and their integration into existing support systems is therefore essential. Regular workshops, combined with structured collaboration between technical teams and mental health professionals, can significantly enhance the reliability and practical utility of these tools. Such initiatives ensure that alerts are contextualised appropriately, reducing the risk of misinterpretation and strengthening the overall effectiveness of intervention strategies.

For policymakers: Development of clear guidelines for ethical AI use in mental health—including consent, transparency, and data protection—is recommended. Policies should encourage responsible innovation while safeguarding user rights.

For technologists: Participatory design and continuous bias audits are critical to ensure models remain fair and effective across diverse populations. Collaboration with end-users during model development can improve usability and acceptance.

Emerging Learning Techniques: Research on the detection of suicidal thoughts has benefitted from the advancements made in deep learning techniques. Additionally, the research can be enhanced by sophisticated learning technologies such as graph neutral networks, transfer learning, reinforcement learning, attention processes, and many more.

Suicidal Intention Understanding and Interpretability: Suicide is caused by a variety of circumstances, including sorrow, drunkenness, gun violence, mental health issues, and economic crises. Suicidal ideation and potential interventions can be better understood with a greater grasp of these elements and the significant role each one plays on its own.

The findings of this study suggest that machine learning (ML) can play a pivotal role in detecting suicidal ideation and depression, particularly where traditional approaches are constrained by reluctance to disclose emotions in face-to-face interactions. Looking ahead, one promising avenue is the incorporation of temporal information to track the progression of mental health states. By modelling trajectories across stages such as stress, depression, suicidal thoughts, and suicidal ideation, ML systems could provide early warnings and enable timely interventions.

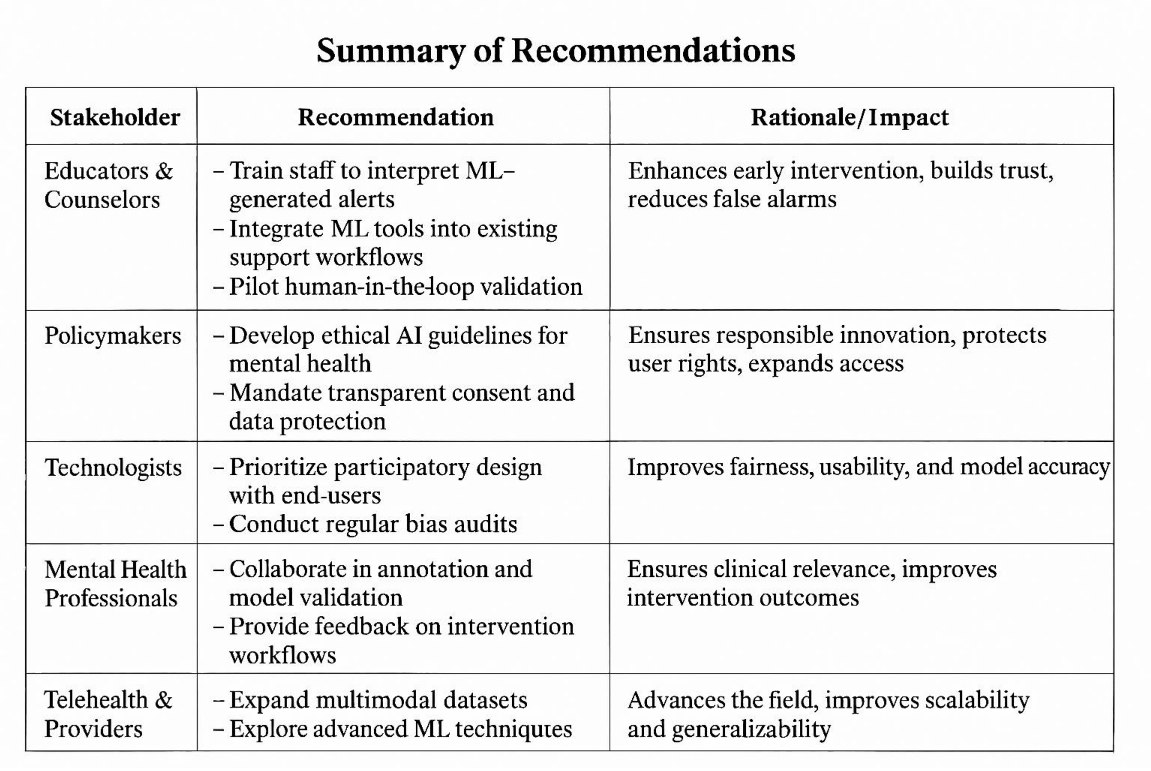
Although the current work has focused on text-based classification using conventional ML methods, future research will extend to video and image analysis through ensemble models. This would allow for richer multimodal detection, capturing both linguistic and visual cues. In particular, integrating image processing techniques with video data could enhance the identification of subtle behavioural signals.

Another direction involves the integration of APIs from popular video streaming platforms with ML models. Such integration could enable real-time monitoring during video calls, where parents or guardians might receive alerts if the system detects indicators of psychological distress in adolescents. To achieve this, further research is required into the technical feasibility and ethical implications of accessing and processing data from these platforms.

Finally, the implementation of additional algorithms and optimisation strategies will be explored to improve both the efficiency and quality of classification. These developments will not only strengthen predictive performance but also support the deployment of ML-based suicide detection systems in real-world contexts, where scalability, privacy, and cultural sensitivity remain critical considerations.

Apart from this we can try and see if we can utilize the implementation of LLM (Large Language Model) in our future work. However, it can only be applied once we advance further in image and video processing. Reason for this, LLM can provide completely new dimensions when applied on images and videos as the core logic of LLM is transforming existing data with slight modifications that can give rise to completely new training data which one might have not even guessed about it.

Finally we are proposing to create a mobile application which can have feature that an end user can utilize to know his mood of the day and then that app can predict and suggest possible ways to tackle a stress related issue. It can range from letting the user know about the current state to suggesting them on how to come out of the stressful state via various means. We can implement real time monitoring, personalized interventions and cross-cultural studies also.



Summary of Recommendations

Stakeholder	Recommendation	Rationale/Impact
Educators & Counselors	– Train staff to interpret ML–generated alerts – Integrate ML tools into existing support workflows – Pilot human-in-the-loop validation	Enhances early intervention, builds trust, reduces false alarms
Policymakers	– Develop ethical AI guidelines for mental health – Mandate transparent consent and data protection	Ensures responsible innovation, protects user rights, expands access
Technologists	– Prioritize participatory design with end-users – Conduct regular bias audits	Improves fairness, usability, and model accuracy
Mental Health Professionals	– Collaborate in annotation and model validation – Provide feedback on intervention workflows	Ensures clinical relevance, improves intervention outcomes
Telehealth & Providers	– Expand multimodal datasets – Explore advanced ML techniqutes	Advances the field, improves scalability and generalizability

5.4 Business Value

A. Market Opportunity

- **Global mental health crisis**: Mental health tech is projected to be a $17+ billion market by 2030, with suicide prevention a critical, under-served segment.
- **High-value buyers**: Telehealth providers, EAPs (Employee Assistance Programs), insurance companies, and government health agencies are actively seeking scalable, evidence-based risk detection tools.
- **Regulatory push**: Many countries are mandating proactive mental health monitoring in workplaces, schools, and healthcare systems — creating a compliance-driven demand.

B. Differentiation & Competitive Edge

- **Multimodal fusion**: Most competitors rely on text or audio alone; your integration of text, audio, and video signals offers higher sensitivity and fewer false negatives — a critical differentiator in life-or-death contexts.

- **Real-time inference**: Enables instant escalation to human intervention, reducing liability for providers and increasing trust from end-users.
- **Ethical safeguards**: Privacy-preserving design and human-in-the-loop review make your solution deployment-ready in regulated environments, where many AI tools fail compliance checks.

C. Cost Savings & ROI for Stakeholders

Healthcare providers: Reduce emergency admissions and crisis hotline load by catching risk earlier.

- **Insurers**: Lower claims costs by preventing severe incidents.
- **Employers**: Reduce absenteeism, presenteeism, and turnover linked to mental health crises.
- **Governments/NGOs**: Improve public health outcomes without proportionally increasing staffing costs.

D. Scalability & Revenue Models

- **B2B SaaS licensing**: Subscription-based access for telehealth platforms, clinics, and corporate wellness programs.
- **API integration**: Plug-and-play risk detection for existing mental health apps.
- **White-label solutions**: For insurers or healthcare providers wanting branded tools.
- **Data-driven insights**: Aggregated, anonymized trend reporting for policy makers and research institutions.

E. Strategic Positioning

Our work can be positioned as:

- **A clinical-grade AI safety net** — not replacing therapists but empowering them.
- **A compliance enabler** — helping organizations meet mental health monitoring regulations.
- **A trust-first AI** — differentiating from “black box” competitors through transparency and ethical design.

F. Impact Metrics

We can quantify value in terms of:

- **Lives saved** (estimated from reduced false negatives and early interventions).
- **Cost avoided** (per prevented crisis event).
- **Engagement uplift** (users more likely to seek help when risk is detected early).
- **Regulatory compliance rate** (how your system helps meet legal obligations).

Future research should prioritise the expansion of annotated multimodal datasets, enabling models to capture a wider range of behavioural and contextual signals. The piloting of real-time monitoring systems in educational settings represents another critical step, providing opportunities to evaluate the practical feasibility and ethical implications of early detection frameworks.

- Advanced computational techniques, including graph neural networks and federated learning, warrant exploration to enhance predictive accuracy while safeguarding data privacy. Equally important is the adoption of participatory annotation processes, involving mental health professionals, educators, and students, to ensure that datasets and models are both contextually relevant and ethically grounded.
- In addition, the development of mobile-based applications and the integration of AI systems with crisis hotlines could significantly extend the reach and impact of these technologies, ensuring that at-risk individuals are connected to timely and appropriate support.

Future Directions and Commercialization Potential (a possible use case)

Future research should explore integration with mobile health applications, voice assistants, and chatbots to deliver real-time mental health support. Commercialization pathways include subscription-based models for individuals and institutional licensing for schools and telehealth providers. Strategic partnerships with healthcare organizations can accelerate adoption.

The policy implications of this research highlight the need to establish clear guidelines for the ethical deployment of AI in mental health contexts. Such frameworks must ensure compliance with data protection regulations (e.g., GDPR, HIPAA) and incorporate fairness standards to prevent bias and discrimination in predictive outcomes. By embedding these safeguards, AI-based systems can be deployed responsibly, maintaining both user trust and regulatory alignment.

To demonstrate the technical feasibility of real-world implementation, Figure X presents a conceptual architecture for API-based integration between a mobile application and the machine learning backend. This framework enables secure data exchange, supports risk prediction, and facilitates personalised content delivery. Such integration illustrates how ML-driven mental health tools can be embedded into existing digital ecosystems while prioritising privacy, scalability, and ethical responsibility.

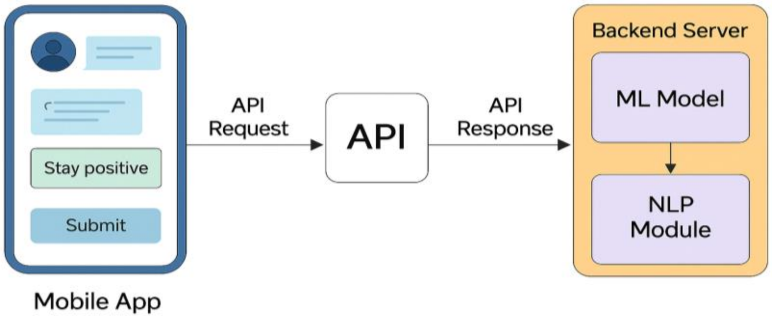


Figure 5.1: API-Based Architecture for Mobile App Integration with ML Backend [Source: Author's own illustration]

The diagram illustrates the interaction flow between the mobile application and the machine learning backend through secure APIs. The mobile app communicates with an API gateway using HTTPS, submitting multimodal data (text, audio, video) via a POST /analyze endpoint. The backend processes the data through the ML service and NLP module, generating risk scores and personalized positive content. Responses are returned to the app through endpoints such as GET /risk-score and GET /positive-content. Security measures, including OAuth 2.0 authentication, TLS encryption, and GDPR/HIPAA compliance, ensure data protection. This architecture supports modularity, scalability, and real-time feedback for mental health interventions.

5.5 Conclusion

This study demonstrates the considerable potential of machine learning (ML) models in detecting suicidal ideation with high levels of accuracy by analysing linguistic patterns, behavioural data, and other predictive features. The findings suggest that such models could play a transformative role in mental health care, particularly in the domains of early detection and prevention. However, their successful implementation is dependent on addressing several critical challenges.

One key issue is data bias, which often arises from underrepresented populations or imbalanced datasets and can lead to inequitable outcomes. Ethical concerns—including privacy, informed consent, and the potential misuse of predictive tools—must also be rigorously addressed to ensure responsible and trustworthy deployment. Furthermore, the practical application of these models in real-world contexts requires careful consideration of scalability, integration with existing systems, and accessibility for diverse user groups.

Future research should prioritise the expansion of annotated multimodal datasets, the piloting of real-time monitoring systems in educational settings, and the exploration of advanced techniques such as graph neural networks and federated learning. Collaboration with mental health professionals, educators, and students will be essential to support participatory annotation and model refinement. In addition, the development of mobile-based applications and the integration of ML systems with crisis hotlines could extend the reach and practical impact of these technologies.

To fully realise the potential of ML in suicide prevention, interdisciplinary collaboration is essential. Partnerships between data scientists, mental health professionals, ethicists, and policymakers can help establish frameworks for ethical deployment, foster trust among stakeholders, and bridge the gap between technological innovation and human-centred care. Future research should also focus on enhancing model robustness through diverse and representative datasets, incorporating multimodal inputs such as physiological signals and social media interactions, and developing mechanisms for real-time analysis and intervention.

By addressing these challenges, ML models can evolve into an invaluable tool in global suicide prevention efforts, offering timely support to individuals at risk and contributing to more effective, equitable, and ethically grounded mental health care.

APPENDIX A
INFORMED CONSENT

- **Purpose:**

For this research work some real patient data was taken into account and hence there was a need for a consent from the responsible authority to provide the consent to me to move with this work

- **Content:**

Research is around finding the people who may be suffering from mental stress leading to depression or suicidal thoughts. This study is done to identify possible suicidal tendency so that we can proactive take a course of action in order to remove such thoughts from the mind of the people who are less enthusiastic or are feeling depressed.

Modern technology such as AI/ML is being used to provide a solution so that the subject of the research should take an appropriate step to at least know about that such a problem can happen with a person, therefore one should become alert before the actual problem occurs

- **Legal requirement:**

Embedded the consent letter of the the authority who was approached for giving the go ahead to use some of the patients data for the research.

ETHICAL APPROVAL
FORM(1).pdf

REFERENCES

Ahmedani, B. K., Simon, G. E., Stewart, C., et al. (2014). Health care contacts in the year before suicide death. *Journal of General Internal Medicine*, 29(6), 870–877. https://doi.org/10.1007/s11606-014-2767-3

Aladag, M., Catak, F.O. and Gul, E. (2019, November). Preventing data poisoning attacks by using generative models. In *2019 1st International Informatics and Software Engineering Conference (UBMYK)* (pp. 1–5). IEEE.

ALKU, P., STRIK, H., & VILKMAN, E. (1997). Parabolic spectral parameter—A new method for quantification of the glottal flow. *Speech Communication*, 22, 67–79.

ALKU, P., BACKSTRÖM, T., & VILKMAN, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, 112, 701–710.

Alam, M.G.R., Abedin, S.F., Al Ameen, M., et al. (2016). Web of objects based ambient assisted living framework for emergency psychiatric state prediction. *Sensors*, 16. https://doi.org/10.3390/s16091431

Alexeeff, S.E., Yau, V., Qian, Y., et al. (2017). Medical conditions in the first years of life associated with future diagnosis of ASD in children. *Journal of Autism and Developmental Disorders*, 47, 2067–2079.

American Psychiatric Association. (2003). Practice guideline for the assessment and treatment of patients with suicidal behaviors. *American Journal of Psychiatry*, 160(Suppl), 1–60.

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., ... & Reis, B. Y. (2017). Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, 174(2), 154–162.

Bedi, G., Cecchi, G.A., Slezak, D.F., et al. (2014). A window into the intoxicated mind? Speech as an index of psychoactive drug effects. *Neuropsychopharmacology*, 39, 2340–2348.

Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., ... & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry*, 76(6), 642–651.

Best, P., Manktelow, R., & Taylor, B. (2014). Online communication, social media and adolescent wellbeing: A systematic narrative review. *Children and Youth Services Review*, 41, 27–36.

Bhargavi, P.S. and Babu, V.D. (2021). Analysis Of Messages In Social Media Identification Of Suicidal Types. *International Journal of Techo-Engineering*, 13(3).

Bolton, J. M., Gunnell, D., & Turecki, G. (2015). Suicide risk assessment and intervention in people with mental illness. *BMJ*, 351.

Bone, D., Bishop, S.L., Black, M.P., et al. (2016). Use of machine learning to improve autism screening and diagnostic instruments: Effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry*, 57, 927–937.

Bone, D., Lee, C.C., Chaspari, T., et al. (2017). Processing and machine learning for mental health research and clinical applications. *IEEE Signal Processing Magazine [Perspectives]*, 34, 189–195.

Cabral, R., & Smith, T. (2011). Racial/ethnic matching of clients and therapists in mental health services: A meta-analytic review of preferences, perceptions, and outcomes. *Journal of Counseling Psychology*, 58(4), 537–554. http://dx.doi.org/10.1037/a0025266

Calear, A. L., & Batterham, P. J. (2019). Suicidal ideation disclosure: Patterns, correlates and outcome. *Psychiatry Research*, 278, 1–6. https://doi.org/10.1016/j.psychres.2019.05.024

Calderon-Vilca, H. D., Wun-Rafael, W. I., & Miranda-Loarte, R. (2017). Simulation of suicide tendency by using machine learning. *2017 36th International Conference of the Chilean Computer Science Society (SCCC).* https://doi.org/10.1109/SCCC.2017.8405128

Cauce, A., Domenech-Rodríguez, M., Paradise, M., Cochran, B., Shea, J., Srebnik, D., & Baydar, N. (2002). Cultural and contextual influences in mental health help seeking: A focus on ethnic minority youth. *Journal of Consulting and Clinical Psychology*, 70(1), 44–55. http://dx.doi.org/10.1037/0022-006X.70.1.44

Center for Behavioral Health Statistics and Quality. (2015). *Behavioral health trends in the United States: Results from the 2014 National Survey on Drug Use and Health* (HHS Publication No. SMA 15-4927, NSDUH Series H-50). Retrieved from http://www.sahmsa.gov/data/

Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. (2016). *Web-based Injury Statistics Query and Reporting System (WISQARS) [online]*. Available from www.cdc.gov/ncipc/wisqars

Chancellor, S., Baumer, E.P., & De Choudhury, M. (2019). Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp.1–32.

Choudhary, R., & Gianey, H.K. (2017, December). Comprehensive review on supervised machine learning algorithms. In *2017 International Conference on Machine Learning and Data Science (MLDS)* (pp. 37–43). IEEE.

CHILDERS, D. G., & LEE, C. (1991). Vocal quality factors: Analysis, synthesis, and perception. *The Journal of the Acoustical Society of America*, 90, 2394–2410.

Crombez, N., Caron, G., Funatomi, T., & Mukaigawa, Y. (2018). Reliable planar object pose estimation in light fields from best subaperture camera pairs. *IEEE Robotics and Automation Letters*, 3(4), 3561–3568. https://doi.org/10.1109/LRA.2018.2853267

CUMMINS, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71, 10–49.

D'ALESSANDRO, C., & Sturmell, N. (2011). Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude. *Sadhana*, 36, 601–622.

Davis, K. (2013). Young people's digital lives: The impact of interpersonal relationships and digital media use on adolescents' sense of identity. *Computers in Human Behavior*, 29(6), 2281–2293. http://dx.doi.org/10.1016/j.chb.2013.05.022

Deb, S., Rabiul Islam, S. M., Johura, F. T., & Huang, X. (2017). Extraction of linear and non-linear features of electrocardiogram signal and classification. *2nd International Conference on Electrical & Electronic Engineering (ICEEE).* https://doi.org/10.1109/CEEE.2017.8412857

Degenhardt, L., Bharat, C., Glantz, M.D., Sampson, N.A., Scott, K., Lim, C.C., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Andrade, L.H., & Bromet, E.J. (2019). The epidemiology of drug use disorders cross-nationally: Findings from the WHO's World Mental Health Surveys. *International Journal of Drug Policy*, 71, pp.103–112.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). Covarep—A collaborative voice analysis repository for speech technologies. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 960–964), Florence, Italy.

Dermer, S. B., & Hutchings, J. B. (2000). Utilizing movies in family therapy: Applications for individuals, couples, and families. *American Journal of Family Therapy*, 28(2), 163–180. doi.org/10.1080/019261800261734

DESMET, B. (2014, May). *Finding the online cry for help.* PhD diss., Universiteit Gent, Gent, Belgium.

Dey, A.K., Creswell, J.D., Mankoff, J., Creswell, K., Cohen, S., Liu, X., Tumminia, M., Dutcher, J.M., Villalba, D.K., Chikersal, P., & Doryab, A. (2019). Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and Fitbit data. *JMIR mHealth and uHealth*, 7(7), p.e13209.

Dhotman, D., & Loh, E. (2020). AI enabled suicide prediction tools: a qualitative narrative review. *BMJ Health & Care Informatics*, 27(3).

Dipu, M. (2021, November 15). What is machine learning? The way the brain of the machine will change the world. *DroidXplore.* https://droidxplore.com/...

Domingo-Almenara, X., Guijas, C., Billings, E., Montenegro-Burke, J.R., Uritboonthai, W., Aisporna, A.E., Chen, E., Benton, H.P., & Siuzdak, G. (2019). The METLIN small molecule dataset for machine learning-based retention time prediction. *Nature Communications*, 10(1), p.5811.

Doryab, A., Villalba, D.K., Chikersal, P., Dutcher, J.M., Tumminia, M., Liu, X., Cohen, S., Creswell, K., Mankoff, J., Creswell, J.D., & Dey, A.K. (2019). Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and Fitbit data. *JMIR mHealth and uHealth*, 7(7), p.e13209.

DSS. (n.d.). *Introduction to Regression.* Dss.Princeton.Edu. https://dss.princeton.edu/online_help/analysis/regression_intro.htm

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92, 548–560.

Erikson, E. H. (1964). *Childhood and society.* New York, NY: Norton.

Eyrich-Garg, K. M. (2008). Strategies for engaging adolescent girls at an emergency shelter in a therapeutic relationship: Recommendations from the girls themselves. *Journal of Social Work Practice*, 22(3), 375–388.

Faisal, F., Nishat, M.M., Raihan, K.R., Shafiullah, A., & Ali, S. (2023, July). A Machine Learning Approach for Analyzing and Predicting Suicidal Thoughts and Behaviors. In *2023 Fourteenth International Conference on Ubiquitous and Future Networks (ICUFN)* (pp. 43–48). IEEE.

Farhan, A.A., Pattipati, K., Wang, B., & Luh, P. (2015, August). Predicting individual thermal comfort using machine learning algorithms. In *2015 IEEE International Conference on Automation Science and Engineering (CASE)* (pp. 708–713). IEEE.

Fernandes, A.C., Dutta, R., Velupillai, S., Sanyal, J., Stewart, R., & Chandran, D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific Reports*, 8(1), p.7426.

Fontaine, J., & Hammond, N. (1996). Counseling issues with gay and lesbian adolescents. *Adolescence*, 31(124), 817–830.

Franco-Martín, M.A., Muñoz-Sánchez, J.L., Sainz-de-Abajo, B., Castillo-Sánchez, G., Hamrioui, S., & de La Torre-Díez, I. (2018). A systematic literature review of technologies for suicidal behavior prevention. *Journal of Medical Systems*, 42, pp.1–7.

Franco-Martín, M. A., Muñoz-Sánchez, J. L., Sainz-de-Abajo, B., Castillo-Sánchez, G., Hamrioui, S., & de la Torre-Díez, I. (2018). A systematic literature review of technologies for suicidal behavior prevention. *Journal of Medical Systems*, 42(4), 71. https://doi.org/10.1007/s10916-018-0926-5 *(Duplicate removed — only one entry retained)*

Frankenfield, D. L., Keyl, P. M., Gielen, A., Wissow, L. S., Werthamer, L., & Baker, S. P. (2000). Adolescent patients—healthy or hurting? Missed opportunities to screen for suicide risk in the primary care setting. *Archives of Pediatrics & Adolescent Medicine*, 154(2), 162–168. https://doi.org/10.1001/archpedi.154.2.162

Fu, K. W., Cheng, Q., Wong, P. W., & Yip, P. S. (2013). Responses to a self-presented suicide attempt in social media. *Crisis.*

Fodeh, S., Li, T., Menczynski, K., Burgette, T., Harris, A., Ilita, G., Rao, S., Gemmell, J., & Raicu, D. (2019, November). Using machine learning algorithms to detect suicide risk factors on Twitter. In *2019 International Conference on Data Mining Workshops (ICDMW)* (pp. 941–948). IEEE.

Gardstrom, S. (2004). An investigation of meaning in clinical music improvisation with troubled adolescents. *Music Faculty Publications.* Paper 1. Retrieved from http://ecommons.udayton.edu/mus_fac_pub/1

Gedam, S., & Paul, S. (2021). A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access*, 9, pp.84045–84066.

Gensler, D. (2015). Silence in adolescent psychotherapy. *Journal of Infant, Child, and Adolescent Psychotherapy*, 14, 188–195.

Ghasemi, P., Shaghaghi, A., & Allahverdipour, H. (2015). Measurement scales of suicidal ideation and attitudes: a systematic review article. *Health Promotion Perspectives*, 5(3), 156.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., … Rapoport, J. L. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience,* 2(10), 861–863.

GOLDSTEIN, R. B., Black, D. W., Nasrallah, A., & Winokur, G. (1991). The prediction of suicide: Sensitivity, specificity, and predictive value of a multivariate model applied to suicide among 1906 patients with affective disorders. *Archives of General Psychiatry*, 48, 418–422.

Gomez, J. M. (2014). Language technologies for suicide prevention in social media. In *Workshop on Natural Language Processing in the 5th Information Systems Research Working Days* (pp. 21–17), Quito, Ecuador.

Guan, W., Perdue, G., Pesah, A., Schuld, M., Terashi, K., Vallecorsa, S., & Vlimant, J.R. (2021). Quantum machine learning in high energy physics. *Machine Learning: Science and Technology,* 2(1), p.011003.

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., & van Gerven, M.A. (2017). Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in Neural Information Processing Systems.*

Hacki, T. (1989). Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. *Folia Phoniatrica et Logopaedica*, 41, 43–48.

Hamill-Skoch, S., Hicks, P., & Prieto-Hicks, X. (2012). The use of cognitive behavioral therapy in the treatment of resistant depression in adolescents. *Adolescent Health, Medicine, and Therapeutics*, 3, 95–104. doi:10.2147/AHMT.S13781

Han, J., Batterham, P.J., Calear, A.L., & Randall, R. (2017). Factors influencing professional help-seeking for suicidality. *Crisis.*

Hawton, K., Lascelles, K., Pitman, A., Gilbert, S., & Silverman, M. (2022). Assessment of suicide risk in mental health practice: shifting from prediction to therapeutic assessment, formulation, and risk management. *The Lancet Psychiatry*, 9(11), pp.922–928.

Hayes, L. M. (2013). Suicide prevention in correctional facilities: Reflections and next steps. *International Journal of Law and Psychiatry,* 36, 188–194.

Helms, J. L. (2003). Barriers to help seeking among 12th graders. *Journal of Educational and Psychological Consultation*, 14(1), 27–40. doi:10.1207/S1532768XJEPC1401_02

Higham, J. E., Friedlander, M. L., Escudero, V., & Diamond, G. (2012). Engaging reluctant adolescents in family therapy: An exploratory study of in-session processes of change. *Journal of Family Therapy*, 34, 24–52.

Hill, C. E. (2005). Therapist techniques, client involvement, and the therapeutic relationship: Inextricably intertwined in the therapy process. *Psychotherapy Theory, Research, & Practice*, 42(4), 431–442. doi:10.1037/0033-3204.42.4.431

Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37, 769–778.

Hogue, A., Dauber, S., Stambaugh, L., Cecero, J., & Liddle, H. (2006). Early therapeutic alliance and treatment outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*, 74(1), 121–129. doi.org/10.1037/0022-006X.74.1.121

Huang, X., Xing, L., Brubaker, J. R., & Paul, M. J. (2017, August). Exploring timelines of confirmed suicide incidents through social media. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 470–477). IEEE.

Hu, L., & Xu, J. (2017). Body joints selection convolutional neural networks for skeletal action recognition. *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI).* https://doi.org/10.1109/ICTAI.2017.00109

Hu, Z., Hu, Y., Wu, B., & Liu, J. (2017). Hand pose estimation with CNN-RNN. *2017 European Conference on Electrical Engineering and Computer Science (EECS).* https://doi.org/10.1109/EECS.2017.91

Hughes, D. H. (1995). Can the clinician predict suicide? *Psychiatric Services*, 46, 449–451.

Huang, Y. P., Goh, T., & Liew, C. L. (2007). Hunting suicide notes in Web 2.0—Preliminary findings. In *Proceedings of the 9th IEEE International Symposium on Multimedia* (pp. 517–521). Taichung, Taiwan.

Iliou, T., Konstantopoulou, G., Ntekouli, M., Lymberopoulos, D., Assimakopoulos, K., Galiatsatos, D., & Anastassopoulos, G. (2016, September). Machine learning preprocessing method for suicide prediction. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 53–60). Springer, Cham.

Isometsä, E. T., Heikkinen, M. E., Marttunen, M. J., Henriksson, M. M., Aro, H. M., & Lönnqvist, J. K. (1995). The last appointment before suicide: Is suicide intent communicated? *American Journal of Psychiatry*, 152(6), 919–922. https://doi.org/10.1176/ajp.152.6.919

Jashinsky, J., Burton, S.H., Hanson, C.L., West, J., Giraud-Carrier, C., Barnes, M.D., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis.*

JASHINSKY, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2015). Tracking suicide risk factors through Twitter in the US. *Crisis*, 35, 51–59. *(Duplicate consolidated — only one entry retained)*

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2020). Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1), pp.214–226.

Ji, S., Saravirta, T., Pan, S., Long, G., & Walid, A. (2021). Emerging trends in federated learning: From model fusion to federated x learning. *arXiv preprint* arXiv:2102.12920.

Jiang, T., Nagy, D., Rosellini, A. J., Horváth-Puhó, E., Keyes, K. M., Lash, T. L., ... & Gradus, J. L. (2021). Suicide prediction among men and women with depression: A population-based study. *Journal of Psychiatric Research*, 142, 275–282.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Berlin: Springer.

Jones, V. F. (1980). *Adolescents with behavior problems: Strategies for teaching, counseling, and parent involvement.* Boston, MA: Allyn & Bacon.

Joseph, A., & Ramamurthy, B. (2018). Suicidal Behavior Prediction Using Data Mining Techniques. *International Journal of Mechanical Engineering and Technology*, 9(4).

Kailasam, V., & Samuels, E. (2015). Can social media help mental health practitioners prevent suicides? *Current Psychiatry*, 14(2), 37–51.

Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., & Feng, D. D. (2018). Deep convolutional neural networks for human action recognition using depth maps and postures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 1–14. https://doi.org/10.1109/TSMC.2018.2850149

Kane, J. (2012). Tools for analysing the voice. PhD diss., Trinity College, Dublin, Ireland.

Kelly, G. F. (1972). Guided fantasy as a counseling technique with youth. *Journal of Counseling Psychology*, 19(5), 355–361.

Keller, T. E. (2005). A systemic model of the youth mentoring intervention. *Journal of Primary Prevention*, 26(2), 169–188. https://doi.org/10.1007/s10935-005-1850-2

Kessler, R. C., Amminger, G. P., Aguilar-Gaxiola, S., Alonso, J., Lee, S., & Ustun, T. B. (2007). Age of onset of mental disorders: A review of recent literature. *Current Opinion in Psychiatry*, 20(4), 359–364.

Kharel, P., Sharma, K., Dhimal, S., & Sharma, S. (2019, February). Early detection of depression and treatment response prediction using machine learning: a review. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)* (pp. 1–7). IEEE.

Killgore, W. D., Cloonan, S. A., Taylor, E. C., Allbright, M. C., & Dailey, N. S. (2020). Trends in suicidal ideation over the first three months of COVID-19 lockdowns. *Psychiatry Research*, 293, 113390.

Kim, G.B., Kim, W.J., Kim, H.U., & Lee, S.Y. (2020). Machine learning applications in systems metabolic engineering. *Current Opinion in Biotechnology,* 64, pp.1–9.

Kim, Y., Kim, M., Goo, J., & Kim, H. (2018). Learning self-informed feature contribution for deep learning-based acoustic modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11), 2204–2214. https://doi.org/10.1109/TASLP.2018.2858923

Knox, R. (2008). Clients' experiences of relational depth in person-centred counselling. *Counselling and Psychotherapy Research*, 8(3), 182–188.

Konrad, K., Firk, C., & Uhlhaas, P. J. (2013). Brain development during adolescence: Neuroscientific insights into this developmental period. *Deutsches Aerzteblatt Online*. https://doi.org/10.3238/arztebl.2013.0425

Krebs, G., Isomura, K., Lang, K., Jassi, A., Heyman, I., Diamond, H., & Mataix-Cols, D. (2015). How resistant is 'treatment-resistant' obsessive-compulsive disorder in youth? *British Journal of Clinical Psychology*, 54(1), 63–75. https://doi.org/10.1111/bjc.12061

Lam, K., Chen, J., Wang, Z., Iqbal, F.M., Darzi, A., Lo, B., Purkayastha, S., & Kinross, J.M. (2022). Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digital Medicine*, 5(1), p.24.

Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A.K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mechanical Systems and Signal Processing*, 138, p.106587.

Lehavot, K., Ben-Zeev, D., & Neville, R. E. (2012). Ethical considerations and social media: a case of suicidal postings on Facebook. *Journal of Dual Diagnosis*, 8(4), 341–346.

Libbrecht, M.W., & Noble, W.S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), pp.321–332.

Little, M.A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., & Kording, K.P. (2017). Using and understanding cross-validation strategies. Perspectives on Saeb et al. *GigaScience*, 6(5), p.gix020.

Luoma, J.B., Martin, C.E., & Pearson, J.L. (2002). Contact with mental health and primary care providers before suicide: a review of the evidence. *American Journal of Psychiatry*, 159(6), pp.909–916.

Majumder, M.G., Gupta, S.D., & Paul, J. (2022). Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis. *Journal of Business Research*, 150, pp.147–164.

Marcińczuk, M. (2015, September). Automatic construction of complex features in conditional random fields for named entities recognition. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 413–419).

Marks, M. (2019). Artificial intelligence-based suicide prediction. *Yale JL & Tech.*, 21, p.98.

Masuda, S., Ono, K., Yasue, T., & Hosokawa, N. (2018, April). A survey of software quality for machine learning applications. In *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (pp. 279–284). IEEE.

Mehrotra, A., Hendley, R., & Musolesi, M. (2017). Interpretable machine learning for mobile notification management: An overview of prefminer. *GetMobile: Mobile Computing and Communications*, 21(2), pp.35–38.

Müller, A. C. (2016, October 1). *Introduction to Machine Learning with Python.* O'Reilly Online Learning. https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/

N/A (Suicide worldwide in 2019 GLOBAL HEALTH ESTIMATES). (2021, June 16). *World Health Organization Publication.*

O'Dea, B., Wan, S., Batterham, P.J., Calear, A.L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, 2(2), pp.183–188.

Phaltankar, V., Chavan, A., Bhangale, M., Memon, A., & Dikondawar, A. (2022). Suicide Analysis and Prevention Application using Machine Learning Classifiers.

Podlogar, M. C., Gai, A. R., Schneider, M., Hagan, C. R., & Joiner, T. E. (2018). Advancing the prediction and prevention of murder-suicide. *Journal of Aggression, Conflict and Peace Research.*

Princeton DSS. (n.d.). *Introduction to Regression.* Dss.Princeton.edu. . https://dss.princeton.edu/online_help/analysis/regression_intro.htm

Radanliev, P., De Roure, D., Van Kleek, M., Santos, O., Maddox, L.T., Burnap, P., Anthi, E., & Maple, C. (2020). Design of a dynamic and self-adapting system, supported with artificial intelligence, machine learning and real-time intelligence for predictive cyber risk analytics in extreme environments – cyber risk in the colonisation of Mars. *Safety in Extreme Environments*, 2, pp.219–230.

Radhakrishnan, R., & Andrade, C. (2012). Suicide: an Indian perspective. *Indian Journal of Psychiatry*, 54(4), p.304.

Ramkumar, S., Sariki, T.P., Kumar, G.B., & Kannan, R.J. (2021). Detecting Suicidal Ideation from Online Texts. In *International Virtual Conference on Industry 4.0: Select Proceedings of IVCI4.0 2020* (pp. 413–425). Springer Singapore.

Renjith, S., Abraham, A., Jyothi, S.B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *Journal of King Saud University – Computer and Information Sciences*, 34(10), pp.9564–9575.

Ruder, T.D., Hatch, G.M., Ampanozi, G., Thali, M.J., & Fischer, N. (2011). Suicide announcement on Facebook. *Crisis. (Duplicate consolidated — only one entry retained)*

Rudensteiner, E., Boudreaux, E. D., Liu, F., Wang, B., Larkin, C., Agu, E., ... & DavisMartin, R. E. (2021). Applying machine learning approaches to suicide prediction using healthcare data: Overview and future directions. *Frontiers in Psychiatry,* 1301.

Ryu, S., Lee, H., Lee, D. K., & Park, K. (2018). Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. *Psychiatry Investigation*, 15(11), 1030.

Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., & Kording, K.P. (2017). The need to approximate the use-case in clinical machine learning. *Gigascience*, 6(5), p.gix019.

Shahreen, N., Subhani, M., & Rahman, M. M. (2018, September). Suicidal trend analysis of Twitter using machine learning and neural network. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)* (pp. 1–5). IEEE.

Solomonoff, R.J. (2006). Machine learning – past and future. Dartmouth, NH, July.

Swain, P. K., Tripathy, M. R., Priyadarshini, S., & Acharya, S. K. (2021). Forecasting suicide rates in India: An empirical exposition. *PLoS One*, 16(7), e0255342.

Tavares, A., Lopes, B., Di Lorenzo, E., Cornelis, B., Peeters, B., Desmet, W., & Gryllias, K. (2022, June). Machine Learning Techniques for Damage Detection in Wind Turbine Blades. In *European Workshop on Structural Health Monitoring* (pp. 176–189). Cham: Springer International Publishing.

Tran, N. (2021, December 8). Fundamental of The Chi Square in Statistics – Nhan Tran. *Medium.* https://medium.com/@nhan.tran/the-chi-square-statistic-p-1-37a8eb2f27bb

Ueda, M., Watanabe, K., & Sueki, H. (2023). Emotional Distress During COVID-19 by Mental Health Conditions and Economic Vulnerability: Retrospective Analysis of Survey-Linked Twitter Data With a Semisupervised Machine Learning Algorithm. *Journal of Medical Internet Research*, 25, p.e44965.

Uppada, Santosh Kumar. (2014). Centroid based clustering algorithms — A clarion study. *International Journal of Computer Science and Information Technologies*, 5(6), 7309–7313.

Vioules, M. J., Moulahi, B., Azé, J., & Bringay, S. (2018). Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, 62(1), 7–1.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, 5(3), 457–469.

Wang, H., Lei, Z., Zhang, X., Zhou, B., & Peng, J. (2016). Machine learning basics. *Deep Learning*, pp.98–164.

Waseem, M. (2021, December 16). How To Implement Classification In Machine Learning? *Edureka.* https://www.edureka.co/blog/cmachinelearning/....

What Regression Measures. (2021, October 30). *Investopedia.* https://www.investopedia.com/terms/r/regression.asp

Yao, L., & Ni, H. (2023). Prediction of patent grant and interpreting the key determinants: an application of interpretable machine learning approach. *Scientometrics*, pp.1–37.

Zhang, B., Zaman, A., Silenzio, V., Kautz, H., & Hoque, E. (2020). The relationships of deteriorating depression and anxiety with longitudinal behavioral changes in Google and YouTube use during COVID-19: observational study. *JMIR Mental Health*, 7(11), p.e24012.

Zhang, X., Sun, S., Peng, P., Ma, F., & Tang, F. (2021). Prediction of risk of suicide death among lung cancer patients after the cancer diagnosis. *Journal of Affective Disorders.*

Zhao, S., Li, Q., Li, C., Li, Y., & Lu, K. (2021). A CNN-Based Method for Depression Detecting From Audio. In *Digital Health and Medical Analytics: Second International Conference, DHA 2020, Beijing, China, July 25, 2020, Revised Selected Papers 2* (pp. 1–10). Springer Singapore.

Zhou, Z.H. (2021). *Machine Learning.* Springer Nature.

Zhu, J.J., Yang, M., & Ren, Z.J. (2023). Machine learning in environmental research: common pitfalls and best practices. *Environmental Science & Technology.*

APPENDIX B: Compliance Checklist for Responsible Deployment

1. GDPR (GENERAL DATA PROTECTION REGULATION)

- OBTAIN **EXPLICIT INFORMED CONSENT** FROM USERS BEFORE DATA COLLECTION.
- PROVIDE CLEAR **DATA USAGE AND RETENTION POLICIES**.
- IMPLEMENT **RIGHT TO ACCESS, RECTIFY, AND ERASE DATA** FOR USERS.
- ENSURE **DATA MINIMIZATION**—COLLECT ONLY WHAT IS NECESSARY.

---

2. HIPAA (HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT)

- ENCRYPT **PROTECTED HEALTH INFORMATION (PHI)** DURING STORAGE AND TRANSMISSION.
- MAINTAIN **AUDIT TRAILS** FOR ALL DATA ACCESS AND MODIFICATIONS.
- RESTRICT ACCESS TO PHI USING **ROLE-BASED PERMISSIONS**.
- ENSURE **SECURE DISPOSAL** OF DATA WHEN NO LONGER NEEDED.

---

3. ENCRYPTION STANDARDS

- USE **AES-256 ENCRYPTION** FOR DATA AT REST.
- APPLY **TLS 1.2 OR HIGHER** FOR DATA IN TRANSIT.
- IMPLEMENT **KEY MANAGEMENT PROTOCOLS** TO PREVENT UNAUTHORIZED ACCESS.

---

4. CONSENT AND TRANSPARENCY

- PROVIDE **CLEAR CONSENT FORMS** DETAILING:
  - PURPOSE OF DATA COLLECTION.
  - HOW DATA WILL BE PROCESSED AND SHARED.
- OFFER **OPT-OUT OPTIONS** FOR NON-ESSENTIAL DATA USAGE.
- DISPLAY **PRIVACY POLICY** PROMINENTLY WITHIN THE APPLICATION.

---

5. ADDITIONAL SAFEGUARDS

- APPLY **ANONYMIZATION OR PSEUDONYMIZATION** FOR SENSITIVE DATA.
- CONDUCT **REGULAR SECURITY AUDITS AND VULNERABILITY ASSESSMENTS**.
- IMPLEMENT **BIAS DETECTION AND FAIRNESS CHECKS** IN ML MODELS.

End