2023 International Research Conference of Sri Lanka Technology Campus
Colombo, Sri Lanka. 14th-15th December 2023

Predictive Modeling for Identifying Insomnia Risk Factors: An Investigative Approach

H.M.S.S. Herath
Computational Intelligence and Robotics Research Lab
Sri Lanka Technological Campus
Padukka, Sri Lanka
sewmih@sltc.ac.lk
https://orcid.org/0009-0008-9702-5576

H.M.K.K.M.B. Herath
Computational Intelligence and Robotics Research Lab
Sri Lanka Technological Campus
Padukka, Sri Lanka
kasunkh@sltc.ac.lk
https://orcid.org/0000-0002-1873-768X

Abstract*—Global populations are significantly impacted by insomnia, a prevalent sleep problem that negatively impacts daily functioning and general well-being. The intricacies of insomnia are explored in this study by utilizing a large dataset that includes both self-reported tests and thorough questionnaires covering various topics, including sleep habits, stress levels, early life events, and cognitive impairments. The study's main objectives are finding relevant components, examining correlations, and utilizing predictive modeling approaches to reveal important insights. We used advanced feature selection techniques to understand the complex interactions between variables. This study examined the intricacies of insomnia's effects on adolescents utilizing a range of statistical metrics, including correlation coefficients and *p*-values. *P*-values, which show how significant the observed links are, and correlation coefficients, which show how strong and which way the relationships are going, are important metrics in our analysis. Using a variety of machine learning methods, such as Decision Trees (DT), k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Naive Bayes (NB), one of the study's main goals was to predict insomnia-related outcomes. Among the models evaluated, the Decision Tree classifier was the most accurate, with an exceptional accuracy rate of 89.47% for both feature selection strategies. These results highlight how reliable Decision Trees are at identifying patterns of sleeplessness. Additionally, the investigation found statistically significant correlations between particular demographic characteristics and insomnia. An important positive link between sex and insomnia was found, with a correlation coefficient of 0.078 and a *p*-value of 0.001. Age and insomnia showed a significant positive link (correlation coefficient = 0.250). However, the *p*-value of 0.553 suggests that more research is needed to understand this relationship fully. Further supporting the need to consider these factors for a thorough understanding and management of insomnia, the study found significant correlations between race (correlation coefficient = 0.05, *p*-value = 0.0) and ethnicity (correlation coefficient = 0.179, *p*-value = 0.716) with insomnia.

Keywords—Insomnia, machine learning, predictive modeling, sleep disorders, sleep patterns

I. INTRODUCTION

An extensive public health concern with far-reaching effects on populations worldwide is insomnia, a common and complex sleep disorder. The detrimental impacts on people's daily functioning and well-being highlight the importance of thoroughly understanding the issue and implementing appropriate management techniques. Many unpleasant symptoms, such as persistently low energy, impaired concentration, irregular appetite, and unsettling mood swings, are commonly experienced by insomniacs. These expressions severely impair people's capacity to function at work and carry out daily tasks, setting off personal and professional difficulties.

The division of insomnia into discrete subtypes, specifically Acute, Chronic, and Comorbid Insomnia, facilitates a sophisticated understanding of the disorder's diverse manifestations and underlying complexity. Acute insomnia is a condition marked by brief episodes of disturbed sleep, usually brought on by acute stressors or environmental changes. It frequently goes away without the need for ongoing medical care. On the other hand, long-lasting chronic insomnia requires specialized and all-encompassing treatment strategies that address the various factors that contribute to its persistence, such as physiological, psychological, and environmental aspects. Comorbid Insomnia highlights the complex interaction between sleep disorders and more general health issues when it coexists with other medical or psychiatric conditions. This highlights the value of integrated care models that address the underlying illness and related sleep disturbances [1, 2]. Fig. 1 depicts the insomnia variation with age in 2022.

The startlingly high incidence of severe insomnia, which impacts roughly 10% of the world's population, highlights the need to give robust and comprehensive sleep health initiatives top priority on both the individual and societal levels. Significantly, the complex relationships that exist between insomnia and several serious health issues, including diabetes, heart disease, obesity, and depression, underscore the complex reciprocal relationship that exists between the quality of one's sleep and one's general health. The need to implement efficient preventive and management strategies is increased due to the interconnectedness of healthcare systems, which strains them and increases the burden on individuals [3].

A comprehensive and holistic approach is necessary to address the multifaceted issues associated with insomnia. This approach should incorporate lifestyle modifications, cognitive and behavioral therapies, tailored therapeutic interventions, and more significant public health initiatives. Stress reduction methods, sleep hygiene education, and encouraging good sleep habits should all be prioritized to significantly reduce insomnia's prevalence and the health hazards that come with it. Furthermore, developing a deeper comprehension of the complex relationships between sleep, mental health, and physical health is essential to developing long-lasting public health policies promoting early detection and intervention techniques. Fig. 2 shows the variation in US adult's experience of insomnia in 2022.

A. *Background and Significance*

Because of the complex interplay of factors such as circadian rhythms, pubertal development, academic pressures, social demands, reduced parental oversight, and excessive use of digital media, adolescents are especially vulnerable to sleep problems. This vulnerable group frequently struggles with sleep deprivation, irregular sleep patterns, and daytime sleepiness (DS). Furthermore, adolescents often face emotional and social challenges, with a significant proportion experiencing depression symptoms. According to studies, one-third of adolescents experience depressive symptoms, and 20% experience major depressive disorder during their lifetime [4]. The presence of insomnia, DS, and depression not only impairs daily functioning but also raises the risk of substance abuse, accidents, and suicidal ideation.

B. *Insomnia as a Global Health Concern*

Machine Learning (ML) appears to be a promising tool for comprehending and treating the complexities of insomnia. Supervised learning in ML allows algorithms to detect patterns in large datasets and classify outcomes based on pre-labeled variables [5]. This study uses data-driven analysis to identify the underlying factors contributing to insomnia. The study aims to extract actionable insights by delving into specialized insomnia datasets. This investigation seeks to bridge the gap between data-driven analysis and practical solutions for people dealing with insomnia. We hope this research will provide hope and tangible support to those suffering from insomnia, improving their overall well-being and quality of life.

C. *Objectives of the Study*

This research is based on analyzing the dataset based on insomnia, which will aim to understand this disorder and improve the well-being of affected individuals. While identifying the contributing factors such as sleep patterns, personality traits, cognitive processes, stressors, coping mechanisms, and childhood experiences. Using statistical analyses to explore patterns and correlations within the dataset and identify relationships between different variables can provide valuable insights into the interplay of various factors contributing to insomnia. Using ML algorithms and statistical modeling techniques to develop predictive models for insomnia, predict the likelihood of insomnia occurrence based on specific variables, enabling early identification and intervention for at-risk individuals.

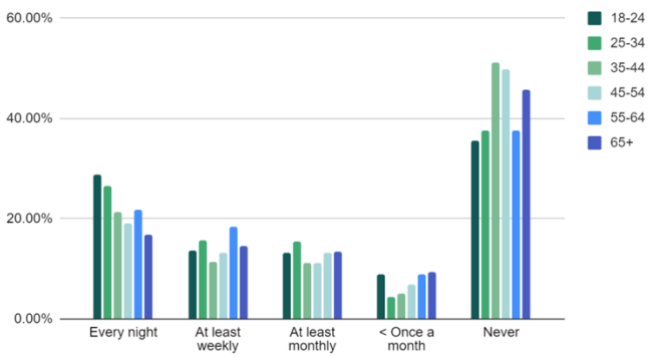

Fig. 1. Frequency of insomnia by age group in 2022, *Adopted from [2]*

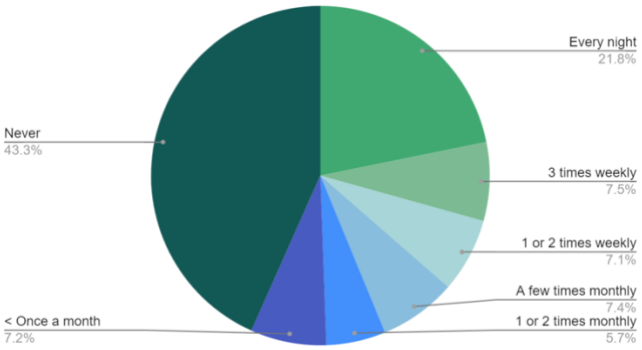

Fig. 2. Occurrence of insomnia among adults in the United States in the year 2022, *Adopted from [2]*

II. Literature Review

The focus of research attention has shifted recently to insomnia, a common sleep disorder, leading researchers to use a variety of approaches, most notably the use of machine learning algorithms. This method has produced insightful findings that clarify the frequency, consequences, and underlying risk factors associated with insomnia. These studies advance our understanding of this disorder and pave the way for novel risk assessment and predictive modeling approaches.

Research findings consistently emphasize the widespread occurrence of insomnia among diverse demographics. A study by researchers [1] delved into the utilization of multiple machine learning algorithms to forecast patterns of insomnia, contributing valuable insights to the expanding literature on insomnia prediction. The use of predictive modeling has dramatically improved the field of health research. Liu et al. [4] conducted a longitudinal study to investigate the relationship between daytime sleepiness, depression symptoms, and insomnia in teenagers. They used predictive modeling techniques to provide crucial new information about the complex relationships between these variables. Their study emphasizes the critical importance of holistic assessments in understanding the varied effects of insomnia and the demand for a thorough comprehension of the complicated dynamics at work. Additionally, Kim's research [5] underscores the changing landscape of insomnia-related technologies, emphasizing the transformative capabilities of artificial intelligence (AI) and machine learning.

Singareddy et al. [6] investigated the relationships between several variables and the emergence of chronic insomnia, including behavioral characteristics, psychiatric and medical disorders, demographics, and polysomnography. The study addressed the shortcomings of earlier research by performing a prospective analysis on a large general population sample over an extended follow-up period. The study used information from the Penn State Sleep Cohort, which included 1395 participants who were followed up after 7.5 years and 1741 adult participants. Those who did not have chronic insomnia at the start of the study (n = 1246) were included. Sleep history was collected at both the baseline and follow-up visits, and baseline assessments included extensive medical and psychiatric histories, personality tests, and an 8-hour polysomnography.

The goal of Inouye et al. [7] was to create and validate a predictive model based on admission characteristics that would anticipate the occurrence of new cases of delirium in

hospitalized elderly medical patients. The study comprised two concurrent prospective cohort studies that were carried out in a university teaching hospital. Among the risk factors that were found, delirium was independently predicted by vision impairment, severe illness, cognitive impairment, and a high blood urea nitrogen/creatinine ratio. A risk stratification system based on these factors successfully identified patients at different risk levels.

In their study of the epidemiology of insomnia, Taylor et al. [8] emphasized the condition's role as a risk factor for several illnesses, with a focus on its predictive relationship to substance abuse, psychological disorders, anxiety disorders, depression, and suicidal thoughts. In addition, sleeplessness was associated with immune system dysfunction, but there was conflicting data about its role in cardiovascular disease and death. On the other hand, it was discovered that the usage of sleeping pills predicted death. Notably, the review noted several shortcomings in the examined studies, such as inadequate control for competing theories and poorly defined criteria for insomnia. Notwithstanding these limitations, the review emphasized the importance of insomnia as a risk factor for deteriorated mental and physical health.

The interdisciplinary nature of studies on insomnia was demonstrated by researchers in [2], who focused on using machine learning techniques to identify potential risk factors associated with insomnia. Huang's work reflects the coming together of technological progress and clinical understanding to produce complex insights into the difficult terrain of insomnia risk factors. Furthermore, Kiss et al.'s extensive dataset [9] provides a complex viewpoint on the symptomatology of teenage insomnia. This dataset, which comes from standardized questionnaires, is helpful for scholars who want to learn more about the complex relationship between insomnia and adolescent mental health. Tab. 1 depicts the summary of the related studies.

TABLE I. SUMMARY OF RELATED WORKS ON INSOMNIA AND INSOMNIA-RELATED MACHINE LEARNING STUDIES

Study	Summary
[1]	They utilized multiple machine learning algorithms to forecast patterns of insomnia, contributing valuable insights to the literature on insomnia prediction.
[3]	They conducted a longitudinal study investigating the relationships between daytime sleepiness, depression symptoms, and insomnia in teenagers, emphasizing holistic assessments for understanding the complex dynamics.
[4]	They highlighted the transformative potential of AI and machine learning in insomnia-related technologies.
[5]	They explored relationships between various factors and chronic insomnia, utilizing a large general population sample and addressing prior research limitations.
[6]	They developed and validated a predictive model for new cases of delirium in elderly medical patients, identifying key risk factors, including vision impairment and cognitive impairment.
[7]	They underscored insomnia's role as a risk factor for various illnesses, particularly highlighting its predictive connections to substance abuse, psychological disorders, anxiety, depression, and suicidal ideation.
[2]	They utilized machine learning techniques to identify potential risk factors associated with insomnia, showcasing an interdisciplinary approach.
[8]	They provided a comprehensive dataset offering insights into teenage insomnia symptomatology, serving as a valuable resource for understanding the complex relationship between insomnia and adolescent mental health.

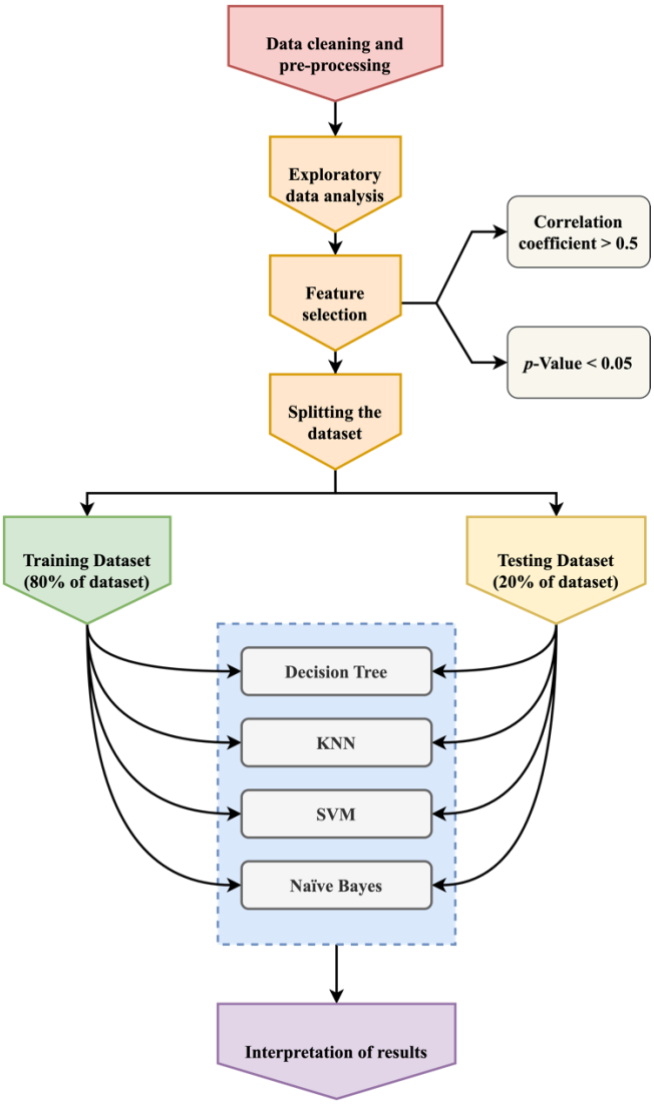

Fig. 3. Model development methodology

III. PREDICTIVE MODEL DEVELOPMENT FOR INSOMNIA RISK ASSESSMENT

In this section, the study's proposed methodology is presented. As depicted in Fig. 3, the model development methodology, aiming to forecast insomnia, is outlined. Preceding the feature selection, the process entails data cleaning and preprocessing.

A. Data Collection and Preprocessing

The dataset used in this study consists of responses from 19 standardized questionnaires and self-reported demographic data, totaling 79 item scores. These questionnaires thoroughly assess a wide range of factors related to insomnia, such as the quality of sleep, routines,

thought patterns, stress levels, coping strategies, emotional control, mood, personality characteristics, and traumatic experiences during childhood. Three separate groups were created from the participant cohort (N = 95): those with sub-clinical insomnia (N = 21, 52% Female), adolescents with clinical insomnia (N = 26, 76% Female), and healthy sleepers (N = 48, 58% Female). To determine insomnia diagnoses, meticulous clinical interviews and adherence to DSM-5 insomnia criteria were used, which is in line with the most recent recommendations for pediatric insomnia [8]. This methodology facilitated a comprehensive and diverse population representation, which is crucial for precise predictive modeling.

Data processing procedures probably included preprocessing activities, including handling missing data, ensuring the data is clean, encoding categorical variables, and standardizing numerical features. This procedure aimed to make the dataset consistent, uniform, and ready for a thorough examination.

Visual and summary methods were used in exploratory data analysis (EDA) to identify patterns, anomalies, and relationships within the dataset. Visualization tools such as correlation matrices and histograms made comprehending the data's fundamental properties and structure easier.

Regarding feature selection—a crucial component in insomnia—the study used methods like correlation analysis and *p*-value computation to determine which features from the questionnaire results are most relevant in influencing insomnia.

Algorithm 1 describes the methodology of the insomnia study analysis.

Algorithm 1 Insomnia Study Analysis

```
Dataset ← LoadDataset
Dataset ← HandleMissingValues(Dataset)
Dataset ← CleanData(Dataset)
Dataset ← EncodeCategoricalVariables(Dataset)
Dataset ← StandardizeNumericalFeatures(Dataset)
VisualizeDataset(Dataset)
SummaryStatistics(Dataset)
RelevantFeatures ← SelectFeatures(Dataset)
Print(RelevantFeatures)
```

B. Selection of Predictive Modeling Techniques

Care must be taken when choosing the proper methods for predictive modeling in assessing insomnia risk. Several machine learning algorithms, such as DT, k-NN, SVM, and NB classifiers, were used. These algorithms were selected based on their ability to identify complex patterns in domains related to insomnia and handle the dataset's complexity. Because decision trees are interpretable, they have been used to specify explicit decision pathways. k-NN, a well-known instance-based learning technique, successfully categorized data points according to their proximity in the feature space.

SVM is a reliable algorithm that has shown usefulness in high-dimensional spaces for regression and classification tasks. NB was chosen as a good probabilistic classifier due to its ease of use and ability to process large datasets. A thorough examination of the dataset and a comprehensive assessment of the insomnia risk factors were made possible by this extensive selection of techniques.

C. Identification of Insomnia Risk Factors

A crucial step in the predictive modeling process was guaranteeing uniformity and consistency in the dataset, which called for extensive data preprocessing. Carefully carried out procedures included handling missing values, cleaning up the data, encoding categorical variables, and standardizing numerical features. Using visualization techniques like correlation matrices and histograms, exploratory data analysis, or EDA, played a crucial role in this process. EDA provided important insights into the structure and characteristics of the dataset by making it easier to identify significant patterns, outliers, and correlations between variables. Feature selection techniques like correlation analysis (see Fig. 4) and *p*-value computations were applied to identify the most relevant features from the questionnaires. The dataset was used to train and assess algorithms during the predictive modeling phase, built upon these particular features.

D. Predictive Modeling and Evaluation

A significant step forward in comprehending and treating insomnia's complexity is applying different machine learning algorithms, such as decision trees, k-NN, SVM, and NB, to predict insomnia-related outcomes based on the chosen features. These advanced techniques made it possible to thoroughly examine the complex interactions between various risk factors and insomnia. Notably, the decision tree algorithm made it easier to identify important pathways in the development of insomnia by offering clear insights into the major factors influencing the condition.

Moreover, proficiently applying k-NN, SVM, and Naïve Bayes algorithms demonstrated the importance of varied analytical techniques in encapsulating the complex characteristics of insomnia risk factors. In summary, this study highlights the potential of predictive modeling to provide critical new understandings of the diagnosis and treatment of insomnia, opening the door to more focused and successful field interventions.

IV. Results and Analysis

This section discusses the results and analysis of the proposed system. Fig. 5 shows the dataset analysis and distribution in various parameters such as gender, age, ethnicity, and Insomnia group.

The diagram in Fig. 6 illustrates the confusion matrix of the feature derived from the correlation coefficient of multiple classifiers. According to the data presented, the DT model demonstrated greater accuracy, yielding higher true positive and false negative outcomes.

The results of the test, derived from the feature's *p*-values, are displayed in Fig. 7. Additionally, the data implies that the DT model exhibited superior accuracy, as evidenced by its higher diagonal values compared to other classifier models.

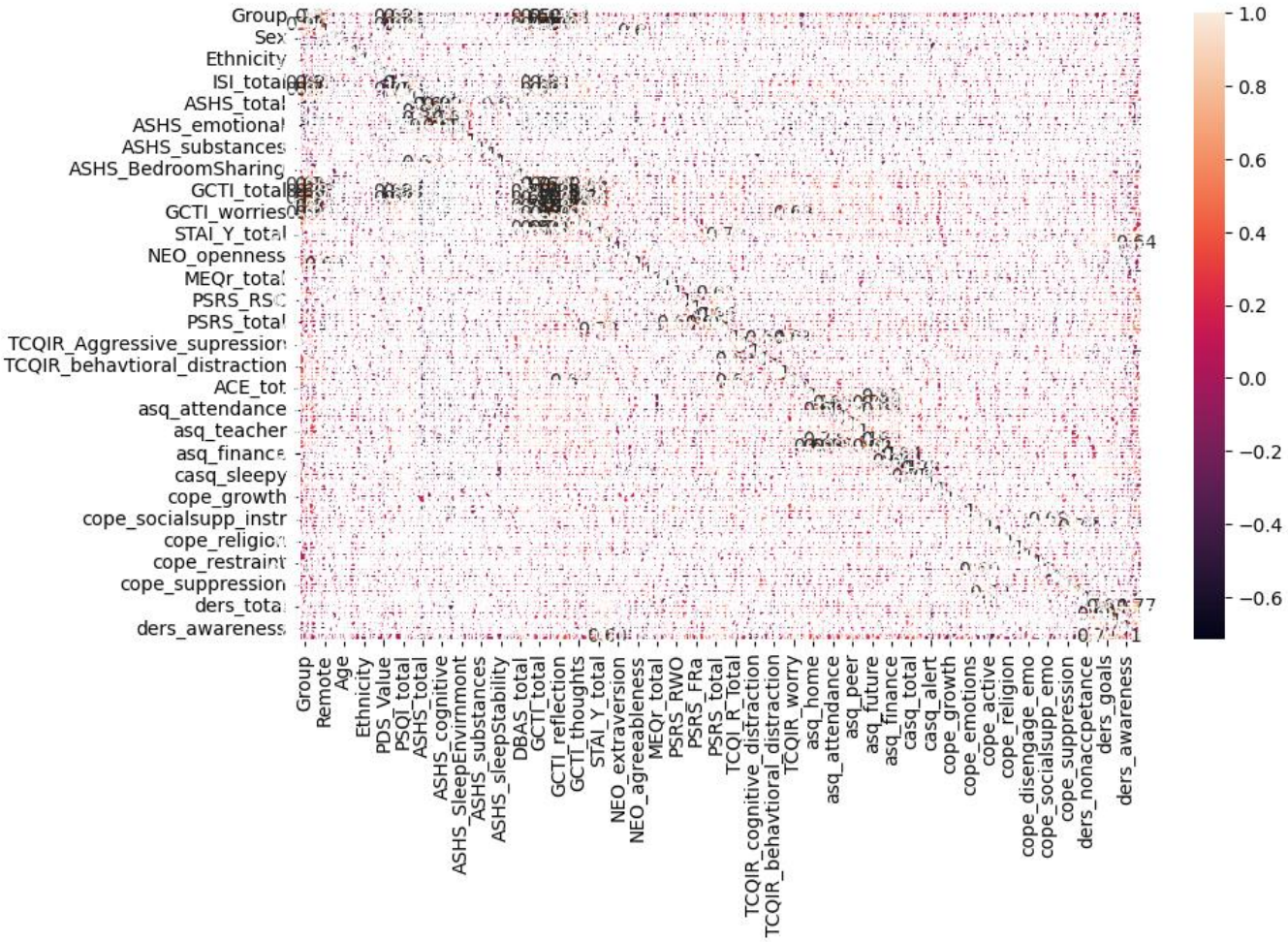

Fig. 4. Correlation coefficient heat map

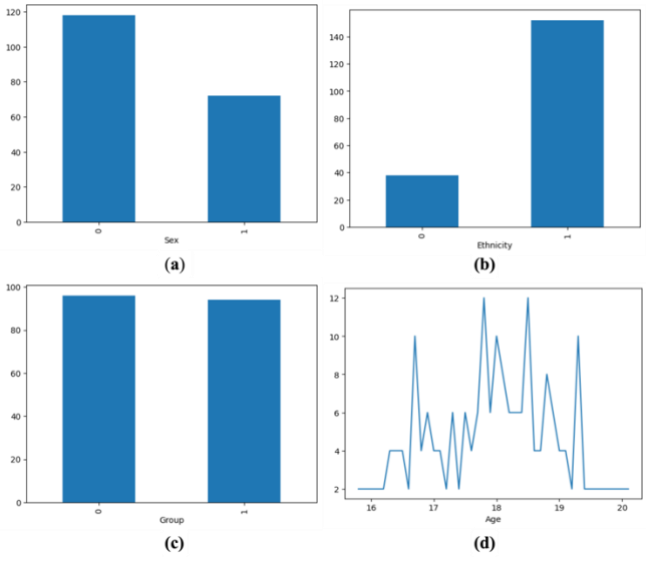

Fig. 5. Data set analysis and distribution for (a) gender, (b) ethnicity, (c) insomnia, and (d) age

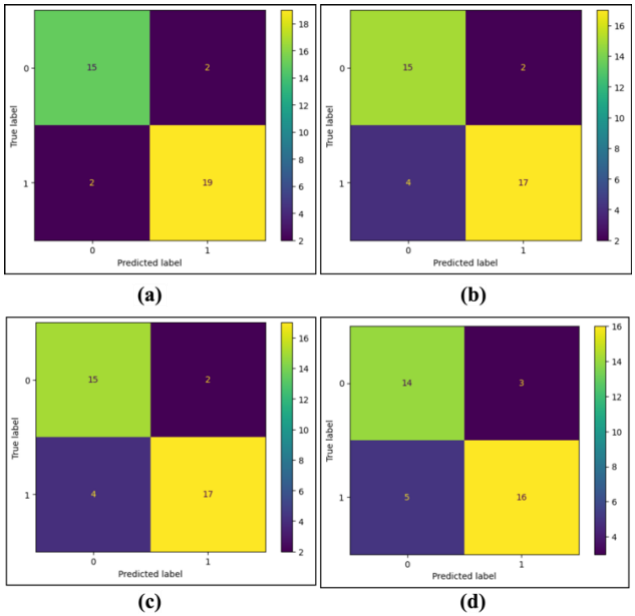

Fig. 6. Confusion matrix of the features extracted from correlation coefficient of (a) DT, (b) k-NN, (c) SVM, and (d) NB

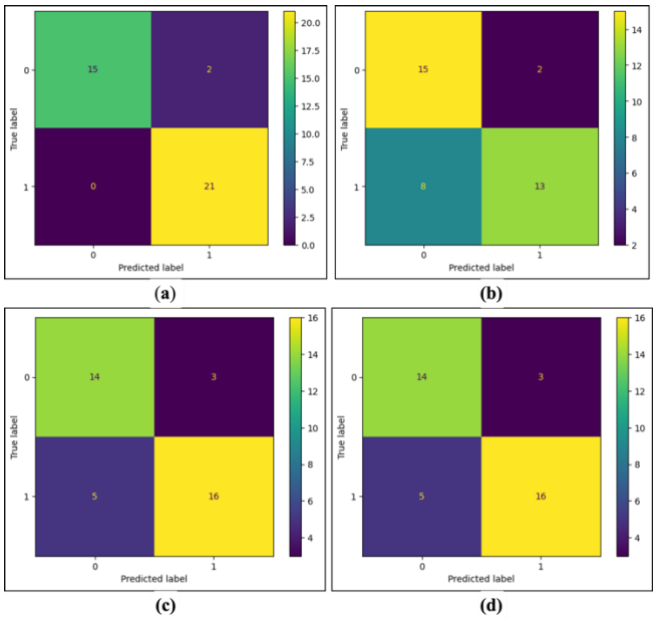

Fig. 7. Confusion matrix of the features extracted from *p*-values of (a) DT, (b) k-NN, (c) SVM, and (d) NB

To investigate the fundamental causes of the different ways in which classifiers with various feature selection techniques perform, more investigation and testing are advised. By gaining a deeper understanding of the complex relationship between feature selection and the outcomes of predictive modeling, such endeavors could lead to the development of more accurate and effective insomnia prediction models.

The non-significant data variables connected to the prediction of insomnia seem to be listed in Tab. 2. The significance of each variable has been assessed using a *p*-value, a metric frequently employed in statistical hypothesis testing. In this case, a *p*-value above a predetermined cutoff point (typically 0.05) suggests that the variable does not have statistical significance when predicting insomnia.

Demographic variables like age and ethnicity are among the first variables examined, producing *p*-values of 0.553 and 0.716, respectively. These findings suggest that for this study, age and ethnicity have no discernible effects on the prediction of insomnia.

Additional examinations cover a variety of behavioral and psychological aspects. For example, the Pittsburgh Sleep Quality Index (PSQI_total) and the Pubertal Development Scale (PDS) have *p*-values of 0.553 and 0.748, respectively, indicating that the information obtained from these measures is not significantly helpful in predicting insomnia. Likewise, evaluations associated with the Athens Insomnia Scale, encompassing factors like sleep environment, daytime sleep, and substance use, exhibit *p*-values between 0.432 and 0.893, signifying an absence of significant correlation with the prognosis of insomnia.

TABLE II. NON-SIGNIFICANCE DATA VARAIBLES OF THE INSOMNIA PREDICTION

Variables	*p*-value
Age	0.553
Ethnicity	0.716
PDS (Pubertal Development Scale_Male/Female)	0.748
PDS_Value (Pubertal Development Scale)	0.72
PSQI_total (PSQI total (Pittsburgh sleep quality index)	0.553
ASHS_SleepEnvirnmont (ASHS Total (Adolescent Sleep Hygiene Scale))	0.893
ASHS_DaytimeSleep (ASHS-Daytime Sleep (Adolescent Sleep Hygiene Scale))	0.109
ASHS_substances (ASHS - Substances (Adolescent Sleep Hygiene Scale))	0.432
GCTI_worries (GCTI General Worries (The Glasgow Content of Thoughts Inventory))	0.365
GCTI_thoughts (GCTI Thoughts About the Environment (The Glasgow Content of Thoughts Inventory))	0.357
PSS_total (PSS Total Score (Perceived stress scale))	0.260
TCQIR_cognitive_distraction (Cognitive Distraction / Suppression TCQI-R (Thought Control Questionnaire Insomnia))	0.054
asq_attendance (Stress of School Attendance ASQ (Adolescent Stress Questionnaire))	0.570
asq_leisure (Stress of School / Leisure Conflict ASQ (Adolescent Stress Questionnaire))	0.430
asq_responsibility (Stress of Emerging Adult Responsibility ASQ (Adolescent Stress Questionnaire))	0.492
cope_disengage_su (Mental Disengagement COPE (Coping skills))	0.811
cope_socialsupp_instr (Seeking Social Support - Instrumental COPE (Coping skills))	0.281
cope_socialsupp_emo (Seeking Social Support - Emotional COPE (Coping skills))	0.080
ders_impulse (DERS Impulse control difficulties (IMPULSE) (Difficulties in Emotion Regulation Scale))	0.958

The Generalized Content Test variables related to thoughts and worries show *p*-values of 0.357 and 0.365, respectively, indicating no significant correlation between the variables and the prediction of insomnia. Additionally, according to the study's parameters, the Perceived Stress Scale (PSS_total) produces a *p*-value of 0.26, suggesting that stress perception may not significantly predict insomnia.

Furthermore, several variables, as assessed by various scales, including the Thought Control Questionnaire-IR, Adolescent Sleep Questionnaire, Coping Responses Inventory, and Difficulties in Emotion Regulation Scale, exhibit *p*-values higher than the traditional significance threshold (0.05). These variables include cognitive distraction, adolescent sleep behaviors, coping mechanisms, social support, and impulse control. These findings point to a lack of meaningful correlations between these behavioral and psychological characteristics and the likelihood of insomnia. The information presented in the table emphasizes how the

variables looked at about insomnia prediction are not significant. However, it's crucial to be aware of the study's limitations and the possibility that more research is required to look into other factors that might be important in predicting insomnia.

The outcomes displayed in Tab. 3 highlight how different predictive models perform when their features are chosen based on correlation coefficients. The DT model consistently demonstrated the highest accuracy, precision, recall, and F1-Score (0.89) among all the classifiers. This shows the DT classifier's robustness and dependability in correctly predicting insomnia-related outcomes based on the chosen features. The precision, recall, F1-Score, and accuracy of the k-NN and SVM models ranged between 0.84 and 0.85, indicating comparable performance levels. This implies that although these models have reasonable predictive power, the DT classifier may be more accurate than them. The NB model is still a good choice for insomnia prediction even though it is marginally less accurate with precision, recall, F1-Score, and accuracy of 0.79. These findings highlight the significance of feature selection and the Decision Tree model's better performance in precisely forecasting patterns of insomnia when compared to the other models under investigation.

TABLE III. EVALUATION OF PREDICTIVE MODELS AFTTER SELECTING FEATURES BY CORRELATION COEFFICIENTS

Classifier	Evaluation Metrics			
	Precision	***Recall***	***F1-Score***	***Accuracy***
Decision tree	0.89	0.89	0.89	0.89
k-NN	0.84	0.85	0.84	0.84
SVM	0.84	0.85	0.84	0.84
Naive Bayes	0.79	0.79	0.79	0.79

The results displayed in Tab. 4 offer an assessment of the predictive models' efficacy after the *p*-value-based feature selection. Interestingly, the DT classifier consistently showed 0.89 accuracy, F1-Score, precision, and recall, indicating its robustness in correctly predicting insomnia-related outcomes. Conversely, the k-NN model performed relatively worse, showing lower recall, accuracy, F1-Score, and precision values at 0.76, 0.75, 0.74, and 0.74, respectively. The SVM and NB models demonstrated a comparatively stable performance, with precision, recall, F1-Score, and accuracy all measuring 0.79. These findings support the significance of feature selection techniques in improving the precision of predictive models for insomnia. While the Decision Tree model maintained high predictive accuracy, the k-NN model demonstrated a decline in predictive performance.

TABLE IV. EVALUATION OF PREDICTIVE MODELS AFTER SELECTING FEATURES BY *P*-VALUE

Classifier	Evaluation Metrics			
	Precision	***Recall***	***F1-Score***	***Accuracy***
Decision Tree	0.89	0.89	0.89	0.89
k-NN	0.76	0.75	0.74	0.74
SVM	0.79	0.79	0.79	0.79
Naive Bayes	0.79	0.79	0.79	0.79

V. DISCUSSION AND CONCLUSION

Several variables, including demographics, influence a complex disorder, insomnia. After examining the connection between demographic factors and insomnia, our research found no evidence of a significant relationship between the prevalence of insomnia and age or race. This result is consistent with earlier studies that indicate insomnia is not age-specific and can occur at any stage of life. Nonetheless, as previous research has shown, our study emphasizes the possible influence of pubertal maturation on the prevalence of insomnia symptoms. This suggests that more investigation is required to understand the relationship between developmental phases and sleep disorders fully.

Predictive modeling requires careful feature selection to identify risk factors linked to insomnia. Two different approaches—the correlation coefficient and the *p*-value—were used in our investigation, leading to the selection 7 and 65 features, respectively. Significantly, both approaches emphasized the significance of several important metrics, such as the Ford Insomnia Response to Stress Test, the Glasgow Content of Thoughts Inventory (GCTI) [10], the Insomnia Severity Index (ISI) [11], the GCTI Sleep Related Anxiety [12], and the GCTI Reflection and Planning [10]. These particular measures provided insightful information about various psychological and cognitive factors contributing to the onset and aggravation of symptoms associated with insomnia.

The measures identified a variety of potential risk factors for insomnia. Since excessive worry and anxiety about sleep frequently prolong insomnia, high scores on the GCTI Sleep Related Anxiety were indicative of people who were likely to develop insomnia. Similarly, high GCTI Reflection and Planning scores suggested an overactive mind marked by constant planning and rumination, making it difficult to fall asleep and stay asleep. The results of the Ford Insomnia Response to Stress Test demonstrated the connection between increased stress reactions and insomnia, with long-term stress being a factor in sleep disorders. Furthermore, a thorough examination of GCTI results revealed particular cognitive patterns linked to insomnia, like negative thinking and cognitive distortions, which worsen sleep disturbances.

The Insomnia Severity Index (ISI), which has higher scores indicating more severe insomnia symptoms and consequent impairments in daily functioning and quality of life, has also come to be recognized as a significant risk factor in and of itself. These results highlight how crucial it is to have a thorough understanding of the psychological and cognitive factors that contribute to insomnia to create customized interventions for those who struggle with sleep disturbances.

Based on feature selection techniques, the performance of several classifiers (see Fig. 6 and Fig. 7), such as DT, k-

NN, SVM, and NB, was assessed in the comparative analysis of predictive modeling approaches. Using correlation coefficients to select features led to consistently high accuracy, F1-score, precision, recall, and accuracy (0.89) for the classifiers. This consistency across metrics and classifiers implied that features selected according to correlation coefficients allowed for precise predictions without appreciable performance differences.

On the other hand, more variable results were obtained with the *p*-value-based feature selection method, especially for the k-NN model, which showed a significant decrease in accuracy, precision, recall, and F1-score (0.76, 0.75, 0.74, and 0.74, respectively). The possible drawbacks of utilizing *p*-values in feature selection were brought to light by this disparity, particularly for some classifiers. The study underlined how crucial feature selection strategies are in determining how predictive classifiers are, highlighting the need for rigorous evaluation of the data type and particular classifier requirements in research and real-world applications.

Age was not found to be a direct risk factor for insomnia; however, the impact of pubertal changes on sleep patterns emphasizes the importance of thorough evaluations that consider developmental milestones and hormonal fluctuations, especially in adolescents. Furthermore, consistent with prior research highlighting the independence of insomnia risk from ethnic backgrounds, our results imply that ethnicity is not a determining factor in the risk of developing insomnia. While social and cultural contexts may indirectly affect sleep habits, our findings highlight the importance of considering various contextual factors when analyzing the connection between ethnicity and insomnia.

Furthermore, we employed various feature selection strategies and predictive modeling techniques to conduct a thorough analysis of insomnia prediction in our study. To create a solid basis for precise predictions, it was essential to carefully preprocess the dataset and identify relevant features using thorough data analysis and feature selection techniques like correlation coefficients and *p*-values.

Evaluation metrics were used to evaluate the performance of various predictive models, such as DT, k-NN, SVM, and NB. These metrics included precision, recall, F1-score, and accuracy. The results showed that the predictive models were consistent when features were chosen using correlation coefficients; high precision, recall, F1-score, and accuracy values were consistently at 0.89. However, the *p*-value-based feature selection approach demonstrated inconsistent results, especially with a significant reduction in k-NN's predictive power. This disparity highlights how intricate feature selection is and how it affects the performance of various classifiers.

This study emphasizes the value of a thoughtful approach to feature selection and the necessity of matching feature selection methods to the needs of various predictive models. Our work advocates for a thorough understanding of the implications of feature selection techniques in creating precise and successful insomnia prediction models, and it acts as a guide for researchers and practitioners. In the context of insomnia research, future studies may explore the underlying mechanisms that underlie the variability in model performance in greater detail, offering additional insights into the optimum feature selection process for robust predictive modeling.

REFERENCES

[1] R. Ahuja, Vivek Vishal, Manika Chandna, S. Virmani, and A. Banga, "Comparative Study of Various Machine Learning Algorithms for Prediction of Insomnia," IGI Global eBooks, pp. 776–799, May 2022.

[2] "Global Insomnia Statistics in 2022 & 2023," Helsestart.no, 2022. https://www.helsestart.no/news/global-insomnia-statistics (accessed Oct. 26, 2023).

[3] A. A. Huang and S. Y. Huang, "Use of machine learning to identify risk factors for insomnia," PLOS ONE, vol. 18, no. 4, pp. e0282622–e0282622, Apr. 2023.

[4] X. Liu, Y. Yang, Z. Liu, and C.-X. Jia, "Associations between Insomnia, Daytime Sleepiness, and Depressive Symptoms in Adolescents: A Three-Wave Longitudinal Study," Journal of Clinical Medicine, vol. 11, no. 23, pp. 6912–6912, Nov. 2022.

[5] K. Seung-Soo, "Recent Trends of Artificial Intelligence and Machine Learning for Insomnia Research," Chronobiology in Medicine, vol. 3, no. 1, pp. 16–19, 2021, Accessed: Oct. 25, 2023.

[6] Ravi Singareddy et al., "Risk factors for incident chronic insomnia: A general population prospective study," Sleep Medicine, vol. 13, no. 4, pp. 346–353, Apr. 2012.

[7] S. K. Inouye, "A predictive model for delirium in hospitalized elderly medical patients based on admission characteristics," Annals of Internal Medicine, vol. 119, no. 6, p. 474, 1993.

[8] D. J. Taylor, K. L. Lichstein, and H. H. Durrence, "Insomnia as a health risk factor," Behav. Sleep Med., vol. 1, no. 4, pp. 227–247, 2003.

[9] O. Kiss, Dilara Yüksel, D. Prouty, F. C. Baker, and Massimiliano de Zambotti, "A dataset reflecting the multidimensionality of insomnia symptomatology in adolescence using standardized questionnaires," Data in Brief, vol. 44, pp. 108523–108523, Oct. 2022.

[10] S. Suh et al., "Cognitions and Insomnia Subgroups," Cognitive Therapy and Research, vol. 36, no. 2, pp. 120–128, Nov. 2011.

[11] C. M. Morin, G. Belleville, L. Bélanger, and H. Ivers, "The Insomnia Severity Index: Psychometric Indicators to Detect Insomnia Cases and Evaluate Treatment Response," Sleep, vol. 34, no. 5, pp. 601–608, May 2011.

[12] Dilara Yüksel, O. Kiss, D. Prouty, F. C. Baker, and Massimiliano de Zambotti, "Clinical characterization of insomnia in adolescents – an integrated approach to psychopathology," Sleep Medicine, vol. 93, pp. 26–38, May 2022.