KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 4, February 2024, Hal 2058-2068
DOI 10.30865/klik.v4i4.1702
https://djournals.com/klik

Evaluating Machine Learning Models for Mental Health Diagnostics: A Comparative Analysis and Visual Insights

Gregorius Airlangga

Engineering Faculty, Information System Study Program, Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
Email: gregorius.airlangga@atmajaya.ac.id
Email Penulis Korespondensi: gregorius.airlangga@atmajaya.ac.id

Abstract−This study addresses the critical challenge of enhancing mental health diagnostics amidst a surge in global mental disorder prevalence. With mental health conditions predicted to become the leading cause of disability by 2030, there is an urgent need for more effective diagnostic methods that transcend the limitations of traditional frameworks, such as subjectivity and clinician bias. Leveraging the capabilities of machine learning (ML) to analyze complex datasets, this research aims to fill the gap in the comparative effectiveness of various ML models, particularly within the context of imbalanced mental health datasets. We systematically evaluated the performance of diverse ML models—including Random Forest, Gradient Boosting, Support Vector Machines, and others—on a rich dataset embodying a wide spectrum of symptoms and diagnoses. Through advanced data preprocessing techniques, such as innovative handling of missing values and categorical encoding, coupled with RandomizedSearchCV for model optimization, we provided a comprehensive analysis of the models' effectiveness. The application of oversampling strategies addressed the challenge of dataset imbalance, ensuring realistic clinical scenario evaluations. The study's findings are presented through detailed model performance metrics and visual analytics, such as symptom distribution visualizations and correlation cluster maps, enhancing interpretability and clinical relevance. The discussion section explores the practical applicability of these findings in clinical settings, acknowledging limitations and outlining future research directions. In conclusion, the study presents a nuanced narrative of ML model selection and performance evaluation complexities. The superior performance of ensemble methods like Random Forest and Gradient Boosting classifiers for certain diagnoses demonstrates the potential of ML in mental health diagnostics. However, the varied performance across models underscores the importance of context-specific model selection, considering the trade-offs between accuracy, interpretability, and computational efficiency. This research contributes significantly to the field of mental health diagnostics by highlighting models with the greatest promise for clinical application and by providing a framework for future advancements integrating ML into mental health diagnostics.

Keywords: Mental Health; Machine Learning; Comparison; Random Forest; SVC

1. INTRODUCTION

The global escalation of mental health disorders represents one of the most pressing challenges in public health, affecting millions worldwide without discrimination [1]–[3]. The World Health Organization (WHO) has reported an alarming increase in the prevalence of mental disorders, predicting that mental health conditions will be the leading cause of disability worldwide by 2030 [4]–[6]. This surge underscores an urgent need for more effective diagnostic and therapeutic approaches that can accommodate the growing demand for mental health services [7]–[9]. Traditional diagnostic frameworks, heavily reliant on subjective assessments and patient self-reporting, are fraught with limitations, including variability in diagnostic criteria, the influence of clinician bias, and the inherent complexity of mental health symptomatology [10]–[12]. These challenges highlight the critical need for innovative solutions that can provide objective, reliable, and scalable diagnostic tools [13]–[15]. Extensive research has been conducted on the application of machine learning (ML) techniques in healthcare, offering promising avenues for revolutionizing the diagnosis and treatment of various conditions, including mental health disorders [16]–[18]. Pioneering studies [19]–[21] have highlighted the potential of ML algorithms to analyze and interpret complex datasets, revealing patterns and insights that can predict health outcomes with remarkable accuracy. Techniques such as Random Forest, Gradient Boosting, Support Vector Machines, and deep learning have demonstrated significant potential in classifying mental health conditions based on clinical and symptomatic data [22]–[24]. However, a thorough review of the literature indicates a notable gap: there is limited research on the comparative effectiveness of a wide range of ML models, especially in the context of the imbalanced datasets typical of mental health diagnosis, where some conditions are significantly less common than others [25]–[27]. The urgency of enhancing mental health diagnostics cannot be overstated, given the profound impact of mental disorders on individuals' well-being, societal productivity, and global healthcare systems [28]–[30]. The state of the art in mental health diagnostics is rapidly evolving towards integrating artificial intelligence (AI) and ML with traditional clinical practice [31]. These technologies offer the potential to transcend the limitations of conventional diagnostics, providing tools that are not only scalable but also capable of delivering objective and nuanced insights into the complex spectrum of mental health conditions [32]. Despite these advancements, the field is still in its nascent stages, with significant room for exploration and development, particularly in optimizing ML models for accuracy, interpretability, and clinical relevance in the face of dataset challenges such as imbalance and heterogeneity [33].This research aims to bridge the existing gap by systematically evaluating and comparing the effectiveness of various ML models in diagnosing mental disorders, utilizing a rich dataset that encapsulates a broad spectrum of symptoms and diagnoses. The study is designed to address critical challenges, including dataset imbalance and the need for robust model optimization strategies, to ascertain the most effective algorithms for mental health diagnostics. By leveraging advanced data preprocessing techniques, model optimization through RandomizedSearchCV, and comprehensive performance evaluation metrics, this research endeavors to highlight the models that offer the greatest promise for clinical application in the diverse and

complex landscape of mental health diagnostics. The literature reveals a critical gap in the comparative analysis of ML models tailored to mental health diagnostics, particularly in addressing the challenge of imbalanced datasets. Many studies focus on individual algorithms or a narrow set of models, often without comprehensive consideration of preprocessing techniques or imbalance correction strategies such as oversampling [26], [34], [35]. Furthermore, there is a paucity of research that holistically examines the entire ML workflow—from data preprocessing and model selection to optimization and evaluation—in the context of mental health, considering the unique challenges posed by the subjective and multifaceted nature of mental disorders. This study makes several pivotal contributions to the field of mental health diagnostics and ML. Firstly, it offers a detailed comparative analysis of a wide array of ML models, employing a methodical approach to data preprocessing, including innovative techniques for handling missing values and encoding categorical features. Secondly, the research addresses the critical issue of dataset imbalance through the application of oversampling strategies, ensuring that model evaluations reflect realistic clinical scenarios. Thirdly, the study advances the field by optimizing model parameters via RandomizedSearchCV, ensuring that the algorithms' performance is rigorously tested and validated. Finally, through the use of visual analytics, including symptom distribution visualizations and correlation cluster maps, the research enhances the interpretability and clinical relevance of the findings, facilitating their application in real-world diagnostic settings.Building on this introduction, the article unfolds as follows: Section II delves deep into the methodology, elucidating the dataset preparation, model selection process, and evaluation criteria. Section III presents a detailed analysis of the results, offering insights into model performance metrics and visual analytics. In addition, we also discuss the implications of these findings, exploring their practical applicability in clinical environments, highlighting limitations, and suggesting directions for future research. The article concludes with Section IV, which synthesizes the study's contributions and reflects on the potential of integrating ML into mental health diagnostics to improve outcomes and healthcare delivery.

2. RESEARCH METHODOLOGY

In the realm of data-driven research, the journey from raw data to insightful conclusions is both intricate and methodical. As presented in the Figure 1, Our study embarks on this path by first laying the groundwork with the Preparation of the Dataset, a crucial step where we meticulously collect, organize, and format our data to build a solid foundation for our analysis. The subsequent phase, Data Preprocessing, is where we delve into the heart of our dataset, cleansing it of imperfections and preparing it for the analytical rigor ahead. This involves techniques such as imputation for missing values, encoding for categorical data, normalization, and class balancing, all aimed at refining the data into a pristine set ready for modeling. Moving forward, the Model Selection & Optimization stage is where we explore the diverse landscape of machine learning algorithms, selecting the most promising candidates and fine-tuning their parameters to tease out the best possible performance. This optimization is not just about achieving high accuracy; it's about ensuring our models can generalize well to new, unseen data. Once our models are honed to near perfection, the Evaluation of Models step allows us to critically assess their performance. Through a battery of metrics such as precision, recall, and the F1-score, we measure the efficacy of each model, seeking a balance between various aspects of predictive power. Finally, the Utilization of Software Tools encapsulates the end-to-end process of applying sophisticated programming libraries and tools. This not only streamlines our workflow but also enhances the robustness and reliability of our results. Python's rich ecosystem, with libraries like Pandas for data manipulation, Scikit-learn for machine learning, and Matplotlib and Seaborn for visualization, empowers our research, providing the computational muscle and visual clarity to forge ahead in our quest for knowledge. Thus, our study's workflow is a testament to the structured yet flexible approach that modern data analysis and machine learning research demand.

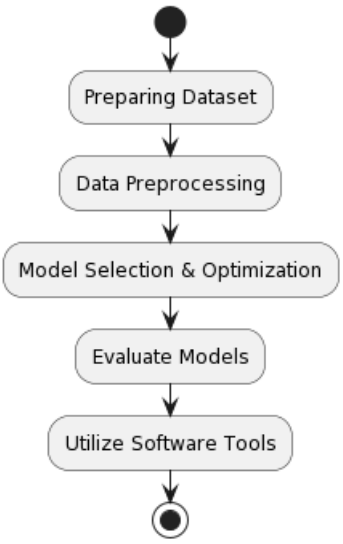

Figure 1. Research Methodology

2.1 Dataset Preparation

The study utilizes a meticulously compiled dataset, "Dataset-Mental-Disorders.csv" and can be downloaded from [36], specifically designed to encapsulate a wide spectrum of mental health disorders and their corresponding symptomatic presentations. This dataset includes a variety of features, such as demographic information (age, gender, etc.), clinical symptoms (mood swings, anxiety levels, thought patterns, etc.), and expert diagnoses. It represents a substantial number of individual cases, each meticulously labeled with a diagnosis determined by mental health professionals. The diversity and depth of the dataset are pivotal, aiming to mirror the real-world complexity of mental health diagnosis, thereby providing a robust foundation for the application and evaluation of machine learning models.

2.2 Dataset Preprocessing

In the initial stages of the research, the dataset was subjected to a thorough cleaning process to eliminate any potential sources of bias, particularly focusing on the removal of irrelevant or identifying columns like 'Patient Number'. This step was crucial to ensure that the subsequent analysis would not be skewed by non-predictive information. Concurrently, the issue of missing values, a frequent challenge in real-world datasets—was addressed. Imputation techniques were applied as presented in the equation (1), filling in missing data points with the mode of their respective columns. This method was selected to preserve the integrity of the dataset and maintain its comprehensiveness, ensuring that the analysis could proceed on a solid foundation of complete data.

$$x_{\text{missing}} = \text{mode}(X) \tag{1}$$

Where x_{missing} represents the missing value within the dataset feature X, and $\text{mode}(X)$ denotes the most frequent value within that feature. Further into the preprocessing phase, the dataset's categorical variables underwent transformation through Label Encoding as presented in the equation (2). This process converted textual labels into a numeric format, rendering the data in a form that machine learning algorithms can process. However, recognizing that raw numerical values could vary widely in scale, a normalization step was also implemented. This step involved scaling numerical values to a common range, a practice especially vital for algorithms like K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), which are sensitive to feature magnitude. Such normalization as presented in the equation (3) ensures that no single feature disproportionately influences the model's outcomes due to its scale, promoting a fair evaluation of all input features.

$$L(c_i) = k \tag{2}$$

Where c_i is the $i-th$ categorical value in the dataset, L denotes the label encoding function, and k is the numeric value assigned to c_i. The dataset also presented a significant challenge in terms of class imbalance, with certain mental health conditions being more frequently represented than others. To mitigate the risk of model bias towards these more common diagnoses, the RandomOverSampler technique was utilized as presented in the equation (4). This approach effectively balanced the class distribution by oversampling the underrepresented classes, ensuring that each condition had an equitable representation in the dataset. This preprocessing step is critical for the success of the study, as it ensures that the machine learning models can learn to identify all conditions with equal accuracy, rather than favoring the more common diagnoses. By addressing these challenges through meticulous data preprocessing, the study set the stage for a fair and comprehensive evaluation of machine learning algorithms in diagnosing mental health conditions.

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)} \times (b - a) + a \tag{3}$$

Where X_{norm} is the normalized value, X is the original value, $\min(X)$ and $\max(X)$ are the minimum and maximum values observed in the dataset, respectively, and a and b represent the scale's desired range, typically 0 and 1.

$$N'_{\text{minority}} = N_{\text{max}} \tag{4}$$

Where N'_{minority} is the new number of instances in each of the minority classes after applying RandomOverSampler, aiming for all classes to have a uniform number of instances N_{max}.

2.3 Model Selection and Optimization

In the pursuit of comprehensively evaluating the potential of machine learning in diagnosing mental health conditions, this study embarked on the selection and optimization of a wide and diverse array of machine learning models, each selected for its unique attributes and the distinct advantages it offers for the classification tasks at hand. The ensemble included models such as RandomForestClassifier, DecisionTreeClassifier. GradientBoostingClassifier, LogisticRegression, SVC (Support Vector Classifier), KNeighborsClassifier, and Naive Bayes as presented in the equation (5) – (10). This variety was intentional, encompassing models ranging from decision trees, which apply hierarchical decision criteria, to SVMs (Support Vector Machines), renowned for their efficacy in delineating the optimal boundary between differing classes. The selection was guided by the principle that different algorithms could offer nuanced insights due to their distinct approaches to handling classification challenges.

$$\text{if } X \leq \text{threshold, go to left child node, else go to right child node} \tag{5}$$

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \tag{6}$$

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{7}$$

$$w^T x + b = 0 \tag{8}$$

$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x_i')^2} \tag{9}$$

$$P(C_k|x_1, \ldots, x_n) \propto P(C_k) \prod_{i=1}^{n} P(x_i|C_k) \tag{10}$$

To refine these models to their most effective forms, the study employed RandomizedSearchCV, a strategic method designed to navigate the vast expanse of possible hyperparameters with which these models can be configured. Unlike traditional grid search methods that methodically explore all possible parameter combinations, RandomizedSearchCV as presented in the equation (15) operates by randomly sampling a subset of parameters for each iteration. This not only expedites the optimization process by evaluating the model's performance across a select number of iterations but also ensures that a broad parameter space is covered, enhancing the likelihood of identifying the most effective model configurations. This optimization strategy is pivotal, as it achieves a delicate equilibrium between computational efficiency and the enhancement of model performance, thereby enabling the selection of hyperparameters that are optimally tuned for the task of diagnosing mental health conditions with the highest possible accuracy and reliability.

2.4 Evaluation Metrics

In The evaluation of machine learning models' performance in this study was meticulously conducted using a suite of key metrics, each offering a unique perspective on the models' diagnostic capabilities. These metrics—accuracy, precision, recall, F1-score as presented in the equation (11)-(14) and the confusion matrix—are essential for a comprehensive assessment, providing insights into not only the overall correctness of the models' predictions but also their ability to accurately identify each class and balance sensitivity with specificity. Accuracy, defined as the ratio of correctly predicted observations (true positives and true negatives) to the total observations, offers a straightforward measure of the models' overall effectiveness. Precision, which measures the proportion of correctly predicted positive observations (true positives) to the total predicted positives (both true positives and false positives), reflects the models' reliability in diagnosing cases as positive. Recall, or sensitivity, indicates the models' ability to identify all actual positives, calculated as the ratio of true positives to the sum of true positives and false negatives. The F1-score, a harmonic mean of precision and recall, provides a single metric to assess the balance between precision and recall, especially useful in situations where an even trade-off between these two metrics is desired. The confusion matrix further complements these metrics by offering a visual and numerical representation of the models' performance, detailing the exact numbers of true positives, true negatives, false positives, and false negatives.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

$$\theta^* = \arg\max_{\theta \in \Theta} f(\theta) \tag{15}$$

2.5 Software Tools

The research extensively utilized the Python programming language, which is highly celebrated for its vast array of libraries and robust community support, particularly in the realms of data science and machine learning. Central to the study's technical framework were several key Python libraries, each selected for its specific capabilities that align with the research's needs. Pandas, a library renowned for its data manipulation and analysis prowess, played a crucial role in handling and preparing the dataset for the subsequent machine learning processes. It enabled the efficient transformation, cleaning, and organization of data, ensuring that the dataset was optimally structured for analysis. Visual representation

of the data and the results of the machine learning models was achieved using Matplotlib and Seaborn. These libraries provided comprehensive data visualization capabilities, allowing for the creation of a variety of plots and graphs that not only facilitated the exploratory data analysis but also enhanced the presentation of the findings, making them accessible and understandable. The visualization tools were instrumental in identifying patterns, outliers, and key insights within the data, as well as in communicating the effectiveness of different machine learning models.

Scikit-learn, another pivotal library in this research, was utilized for the implementation and evaluation of the machine learning models. Known for its wide range of algorithms and tools for model selection, training, and evaluation, Scikit-learn enabled the seamless application of various machine learning techniques, from preprocessing and model selection to hyperparameter optimization and model evaluation. Its comprehensive suite of functionalities supported the rigorous analysis and comparison of the models' performance based on the evaluation metrics. Furthermore, the study addressed the critical issue of class imbalance in the dataset using the Imbalanced-learn library. This library provided tools and techniques specifically designed to handle imbalanced data, allowing for the application of oversampling methods and other strategies to ensure a balanced representation of all classes in the dataset. By leveraging Imbalanced-learn, the research could mitigate the bias towards more prevalent classes, thereby enhancing the models' ability to accurately diagnose a wide range of mental health conditions. Together, these software tools and libraries formed the backbone of the research methodology, enabling the sophisticated analysis and evaluation of machine learning models for diagnosing mental disorders. Their selection was guided by the specific needs of the research, from data preparation and visualization to model implementation and evaluation, ensuring that the study was conducted with the highest standards of accuracy and efficiency.

3. RESULT AND DISCUSSION

3.1 Result Explanation

In our study, we employed a rigorous machine learning methodology to tackle the challenge of diagnosing mental health conditions. Initially, we prepared our dataset by performing data cleaning, which included removing irrelevant features and addressing missing values through mode imputation. We then applied Label Encoding to transform categorical variables into a machine-readable format and normalized the numerical values to ensure uniformity in scale, crucial for the performance of certain algorithms. We selected a suite of machine learning models for their unique capabilities in pattern recognition and classification tasks. Specifically, we utilized Random Forest and Gradient Boosting for their ensemble approaches, which aggregate decisions from multiple trees to improve prediction accuracy and generalizability. Logistic Regression was chosen for its simplicity and interpretability, while the Support Vector Classifier (SVC) was included for its effectiveness in high-dimensional spaces. K-Nearest Neighbors (KNN) was tested for its instance-based learning, which could be advantageous given certain feature distributions. Decision Trees were selected for their hierarchical, rule-based classification, and Naïve Bayes for its probabilistic approach, especially useful when the assumption of feature independence is approximately met. Each model underwent hyperparameter tuning and was trained using a robust cross-validation framework to ensure that our results would be reliable and indicative of each algorithm's performance on unseen data. We evaluated our models using a range of metrics, including accuracy, precision, recall, and the F1-Score, to capture both the models' ability to correctly predict positive instances and their overall reliability.

The comparison of machine learning models in the table 1 shows that Random Forest and Gradient Boosting algorithms both achieve an accuracy and recall of 0.9167, with Random Forest having a slightly lower precision of 0.9340 compared to Gradient Boosting's 0.95, which suggests Gradient Boosting is slightly more reliable when it predicts a positive class. The F1-Score, which balances the precision and recall, is nearly identical for both, with Random Forest at 0.9177 and Gradient Boosting at 0.9132, indicating robust overall performance for both models. Logistic Regression, on the other hand, presents lower scores across the board, with an accuracy of 0.75, precision of 0.8135, recall also at 0.75, and an F1-Score of 0.7547, suggesting that this model may be less effective at this task or require more feature engineering or parameter tuning.

The Support Vector Classifier (SVC) stands out with a high precision of 0.9092 but a lower recall of 0.875, reflected in a fairly high F1-Score of 0.8782, indicating it is quite selective when predicting a positive class but may miss some positive instances. K-Nearest Neighbors (KNN) struggles comparatively, with the lowest scores in all metrics: an accuracy of 0.625, precision of 0.6345, recall of 0.625, and an F1-Score of 0.6163, suggesting that the default parameters or the distance metric may not be well-suited to the dataset. Decision Tree classifiers demonstrate good performance with an accuracy of 0.875, precision of 0.9092, and recall equal to the accuracy, but the F1-Score at 0.8782 indicates there may be room for improvement, perhaps due to a tendency to overfit.

Lastly, the Naïve Bayes model, with an accuracy of 0.833, precision of 0.7651, and recall the same as its accuracy, ends up with an F1-Score of 0.7864, which suggests it is reasonably good at identifying positive instances but not as precise as other models, possibly due to its assumption of feature independence which may not hold in this context. In this study, the choice between these models would be dictated by the particular cost of misclassification specific to the application; for instance, in a medical setting where missing a diagnosis could be dangerous, a model with the highest recall might be preferred. Conversely, in an area where false alarms are costly, a model with higher precision would be more suitable. The F1-Score is particularly useful when looking for a balance between precision and recall, which in this case, points towards ensemble methods as the most effective approach.

Table 1. Comparison Results

Name	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9167	0.9340	0.9167	0.9177
Gradient Boosting	0.9167	0.95	0.9167	0.9132
Logistic Regression	0.75	0.8135	0.75	0.7547
SVC	0.875	0.9092	0.875	0.8782
KNN	0.625	0.6345	0.625	0.6163
Decision Tree	0.875	0.9092	0.875	0.8782
Naïve Bayes	0.833	0.7651	0.833	0.7864

3.2 Discussion

Figure 2 provides a rich visual representation of the relationship between various symptoms and mental health diagnoses as determined by experts. Each panel in the array of histograms with overlaid density plots is a detailed snapshot of one symptom's distribution and how frequently it is associated with each of the diagnoses, which are encoded in colors ranging from 0 to 3. These colors allow for immediate visual segmentation of the data according to the diagnosed condition, presenting a multidimensional view of the symptomatology landscape. As we delve into the interpretation of these histograms, we focus on the x-axis, which quantifies the intensity or frequency of symptoms reported or observed, while the y-axis captures the count of instances within each diagnostic category. The histograms serve not just as a count of occurrences but also as a nuanced depiction of symptom prevalence across different mental health conditions. For instance, a symptom like "Euphoria" might show varied distribution, with certain diagnostic categories exhibiting a higher frequency at increased intensities of this symptom. This could suggest that such a symptom is particularly salient for those categories and might be a significant factor in the diagnostic process.

The correlation between symptom expression and expert diagnosis is revealed through the concentration of colors within specific ranges on the histograms. For example, if one color dominates the higher end of the "Exhaustion" symptom histogram, it indicates a strong association of that symptom with a particular diagnosis. This visual cue helps in discerning which symptoms are indicative of certain conditions and might also highlight symptoms that are common across multiple diagnoses, implying a shared pathology or symptom overlap. Overlaying the histograms are density plots that provide a smooth estimation of the data's distribution, giving a clearer picture of where the majority of data points lie and highlighting the most common symptom intensities for each diagnosis. These curves are particularly useful in identifying the central tendencies and the spread of data, which are critical for understanding the typical and atypical presentations of symptoms.

A comparative analysis across these plots reveals symptom specificity or generality. For instance, symptoms such as "Suicidal Thoughts" might be tightly clustered within a single diagnostic category, suggesting a strong association with that condition. Conversely, a symptom like "Sleep Disorder," if spread across several diagnostic categories, may indicate a common symptom that is not distinct to any single condition. The histograms' shape can also suggest the skewness of the data. A tail extending towards higher symptom intensities might indicate a skew towards more severe symptom manifestation. Additionally, the spread of the data within each histogram can offer insights into the variability of symptom expression. Wide bins with high counts are indicative of a broad range of symptom intensities within certain diagnoses, suggesting heterogeneity in how patients experience and report symptoms.

Outliers or anomalies, if present as separate bins detached from the main distribution, can signal atypical symptom presentations. These could be the result of various factors, including measurement errors, unique patient experiences, or misdiagnoses, and warrant further investigation to understand their origins and implications. For clinicians, the detailed visual information provided by these plots can enhance diagnostic accuracy by highlighting symptom patterns that are strongly associated with particular mental health conditions. Additionally, the overlap of symptoms across different diagnoses revealed in these plots can be essential for recognizing comorbidities. From a research standpoint, these distributions are invaluable for the development of treatment protocols. Symptoms that show a strong correlation with certain diagnoses can be targeted more precisely, leading to more personalized and effective treatment plans.

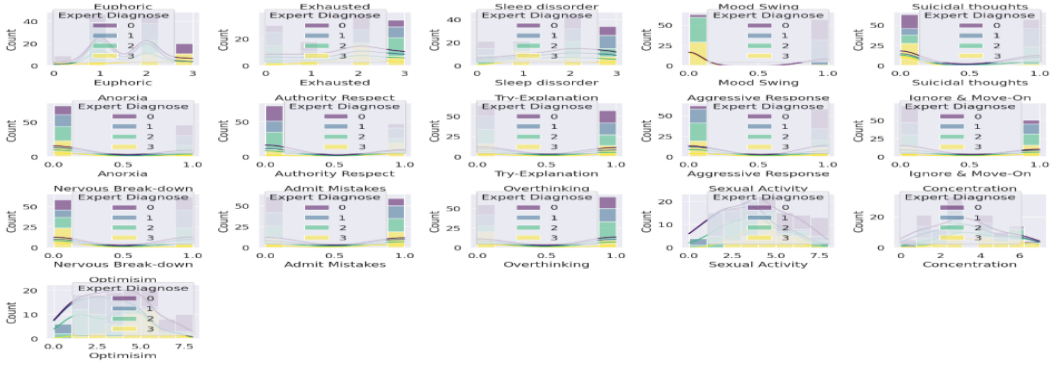

Figure 2. Histogram of Mentality Disorder Symptoms

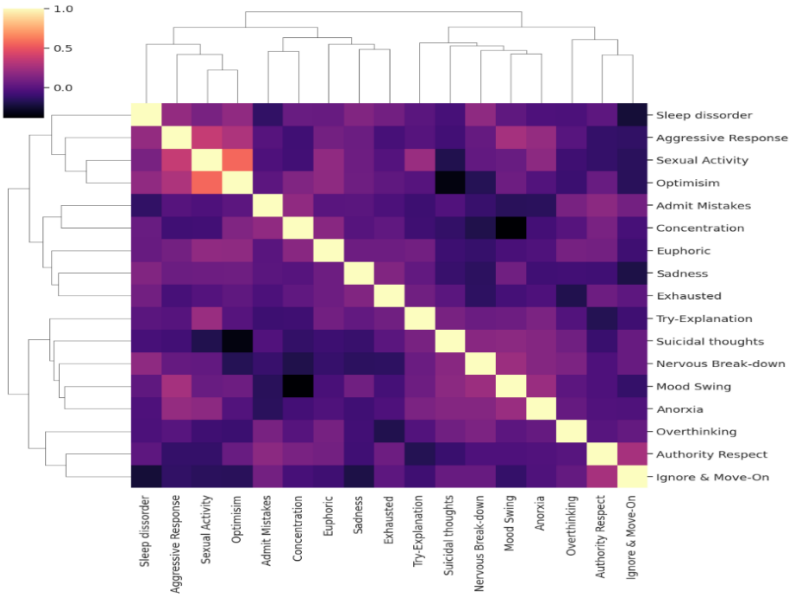

Figure 3. Heatmap of hierarchical clustering

The image in figure 3, is a heatmap complete with hierarchical clustering, serves as a sophisticated visual tool for elucidating the intricate relationships among a range of symptoms commonly encountered in mental health evaluations. This type of visualization is particularly adept at conveying the degree of correlation between variables through a gradation of colors, where each shade corresponds to the strength and direction of the association between symptom pairs. In the context of this heatmap, shades closer to yellow suggest a strong positive correlation, implying that the presence or severity of one symptom is likely to be mirrored by the presence or severity of another. Conversely, darker hues intimate a negative correlation, indicating an inverse relationship between symptoms. The rows and columns, adorned with labels such as "Sleep Disorder," "Euphoric," or "Aggressive Response," represent different psychological symptoms or behaviors. The clustering trees (dendrograms) branching out on both axes group together those symptoms that share similar profiles of correlation, which could be reflective of common underlying factors or indicative of shared pathways in various psychiatric conditions. These dendrograms are not merely ornamental; they convey a hierarchy of associations, revealing how closely or distantly related different symptoms are in the context of the data collected.

Interpreting the heatmap involves an assessment of not only the individual colors but also the broader patterns they form. For instance, if symptoms like "Sadness" and "Exhaustion" are illuminated in similar shades of light color, it could suggest that these symptoms often coalesce, perhaps delineating a specific mental health condition or a symptom cluster. This visualization can thus provide clinicians with a window into the symptomatology of the disorders they treat, highlighting potential areas of focus for diagnostic or therapeutic interventions. Moreover, the heatmap extends beyond clinical utility; it is a trove of insight for research into mental health. The clustering and correlation patterns can guide hypothesis generation for further empirical investigation. For example, if "Optimism" consistently appears in stark contrast to other symptoms, it might prompt researchers to explore its potential as a protective factor against various mental health conditions. However, the heatmap's illustrative power must be contextualized with an understanding of its limitations. While it adeptly maps out correlations, it does not chart causal pathways. The presence of a strong correlation between two symptoms does not delineate whether one causes the other or if they are both manifestations of an underlying process. Additionally, the data encapsulated in the heatmap is but a snapshot, possibly contingent upon the specific sample from which it was drawn and may not necessarily be generalizable across different populations.

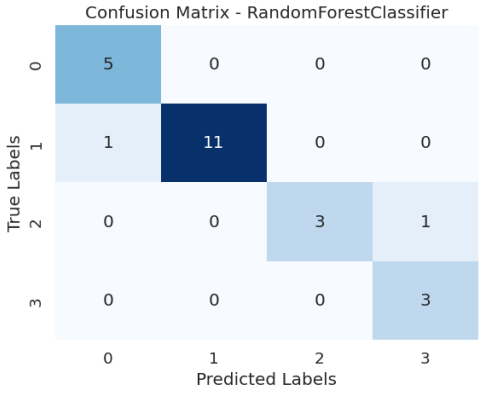

Figure 4. Confusion Matrix of Random Forest Classifier

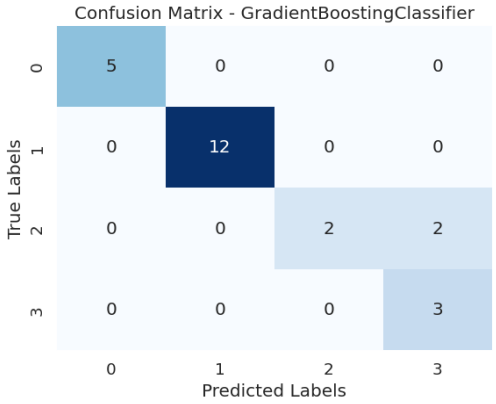

Figure 5. Confusion Matrix of Gradient Boosting Classifier

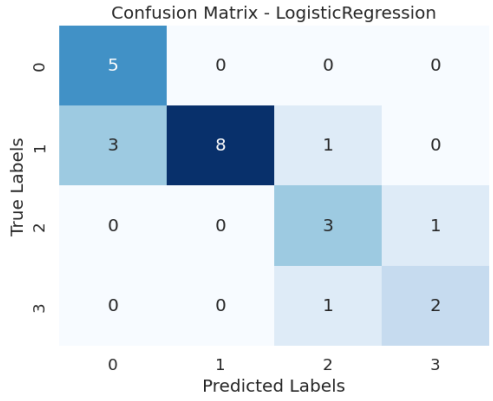

Figure 6. Confusion Matrix of Logistic Regression

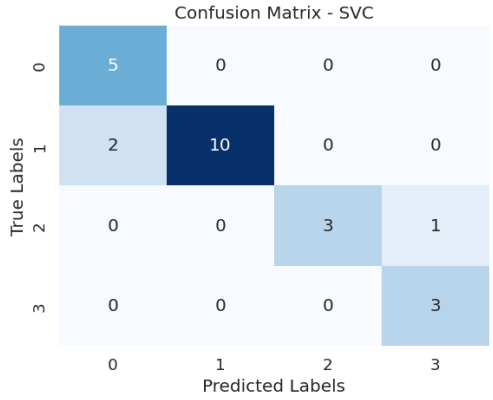

Figure 7. Confusion Matrix of SVC

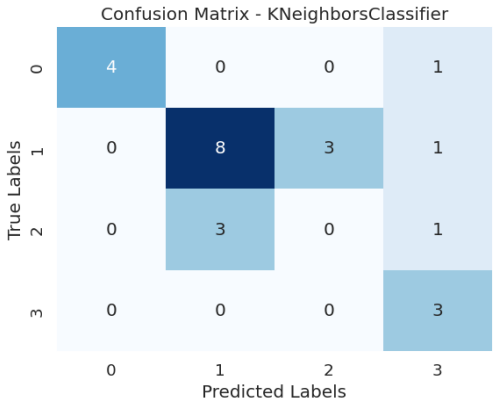

Figure 8. Confusion Matrix of KNN

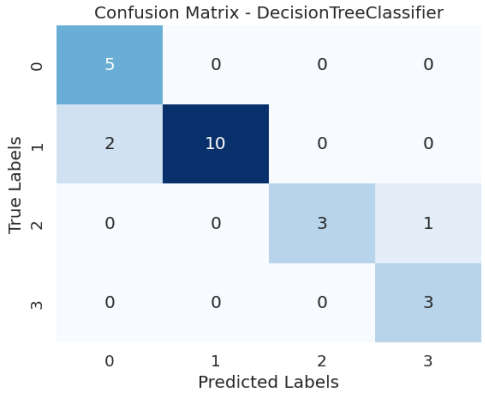

Figure 9. Confusion Matrix of Decision Tree

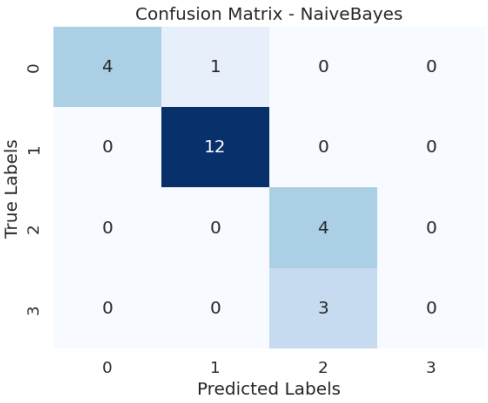

Figure 10. Confusion Matrix of Naïve Bayes

As presented in the figure 4-10 about confusion matrix, The Random Forest Classifier's confusion matrix presents a strong performance, especially in classifying Class 1 with 11 instances correctly identified and only one misclassified. This suggests a robust ensemble method where multiple decision trees manage to create an effective consensus for this class. However, the model exhibits some confusion between Classes 2 and 3, as well as between Classes 0 and 1, suggesting that the random selection of features at each split might sometimes lead to less accurate classification for these particular instances. The Gradient Boosting Classifier shows a similar strength in identifying Class 1 without any misclassification, which indicates that the sequential correction of errors by this model is particularly effective for this class. However, it struggles with Class 2, with misclassifications pointing towards Class 1 and 3. This could be due to the model overfitting on Class 1 features or not being able to capture the complexity of Class 2's feature space sufficiently.

Moving on to the Logistic Regression matrix, the model shows a balanced classification with a few errors across the board. It manages to classify Class 0 perfectly but confuses several instances of Class 1 with Class 0 and Class 2. This might be indicative of the inherent limitation of logistic regression in dealing with non-linearly separable data or overlapping class distributions. The Support Vector Classifier (SVC) offers a decent classification accuracy, particularly for Class 1, similar to the performance of the Gradient Boosting Classifier. However, it misclassifies Class 2 instances as belonging to Class 1 and 3. The SVC's ability to find a hyperplane in a higher-dimensional space through the kernel trick doesn't seem to perfectly separate all classes, which may suggest that the selected kernel is not ideally suited to the underlying data distribution. The K-Neighbors Classifier confusion matrix reveals a model that performs reasonably well for Class 0 but shows significant misclassification for Class 1, where instances are classified as either Class 2 or 3. The model also misclassifies Class 2 instances as Class 1. This indicates that the chosen 'k' value or the distance metric does not perfectly encapsulate the true structure of the data, leading to a reliance on neighboring points that do not necessarily represent the correct class.

The Decision Tree Classifier's performance is quite strong for Class 0, with all instances classified correctly. However, for Class 1, the model confuses it with Class 0, and for Class 2, the model spreads its predictions across Classes 1 and 3. This can imply that the tree might not be deep enough or is over-pruned, leading to insufficient capture of the complexities between the classes. Finally, the Naive Bayes classifier demonstrates a very high precision for Class 1 with complete accuracy and also performs perfectly for Classes 2 and 3. The single misclassification from Class 0 into Class 1 is the only blemish, indicating a slight overlap in the feature distributions assumed by the model. Naive Bayes assumes feature independence, which might not hold true in all cases, leading to potential misclassifications.

4. CONCLUSION

The study's comprehensive evaluation of classifiers reveals the nuanced efficacy of machine learning models, underscoring the importance of matching model strengths to task-specific requirements. Ensemble methods, particularly Random Forest and Gradient Boosting, stand out for their robust performance, affirming their suitability for complex classification challenges where feature interplay is significant. The analysis not only reinforces the value of these methods but also highlights the broader imperative: selecting a machine learning model demands a holistic view of the problem at hand, factoring in the intricacies of data, the cost implications of predictive errors, and the trade-offs between interpretability and computational demand. The goal is to align model capabilities with the specific contours of the application's landscape, ensuring the chosen model delivers reliable insights within the practical confines of its deployment.

REFERENCES

[1] W. Bai *et al.*, "A joint international collaboration to address the inevitable mental health crisis in Ukraine," *Nat. Med.*, vol. 28, no. 6, pp. 1103–1104, 2022.

[2] S. Shoib *et al.*, "Suicide, stigma and COVID-19: a call for action from low and middle income countries," *Front. psychiatry*, vol. 13, p. 894524, 2022.

[3] A. Mallard, M. A. Pesantes, C. Zavaleta-Cortijo, and J. Ward, "An urgent call to collect data related to COVID-19 and Indigenous populations globally," *BMJ Glob. Heal.*, vol. 6, no. 3, 2021.

[4] J. Piao *et al.*, "Alarming changes in the global burden of mental disorders in children and adolescents from 1990 to 2019: a systematic analysis for the Global Burden of Disease study," *Eur. Child \& Adolesc. Psychiatry*, vol. 31, no. 11, pp. 1827–1845, 2022.

[5] I. F. Tso and S. Park, "Alarming levels of psychiatric symptoms and the role of loneliness during the COVID-19 epidemic: A case study of Hong Kong," *Psychiatry Res.*, vol. 293, p. 113423, 2020.

[6] I. Ameer, M. Arif, G. Sidorov, H. Gòmez-Adorno, and A. Gelbukh, "Mental illness classification on social media texts using deep learning and transfer learning," *arXiv Prepr. arXiv2207.01012*, 2022.

[7] S. Gorrell, E. E. Reilly, L. Brosof, and D. Le Grange, "Use of telehealth in the management of adolescent eating disorders: patient perspectives and future directions suggested from the COVID-19 pandemic," *Adolesc. Health. Med. Ther.*, pp. 45–53, 2022.

[8] D. Giansanti, "An Umbrella Review of the Fusion of fMRI and AI in Autism," *Diagnostics*, vol. 13, no. 23, p. 3552, 2023.

[9] V. M. Saceleanu *et al.*, "Integrative approaches in acute ischemic stroke: from symptom recognition to future innovations," *Biomedicines*, vol. 11, no. 10, p. 2617, 2023.

[10] B. N. E. Quinn, "An analysis of the acceptability, feasibility, and utility of the Global Mental Health Assessment Tool for Primary Care (GMHAT/PC) in a UK primary healthcare setting: a practice-based mixed methods study," 2021.

[11] E. Freeman *et al.*, "Patient-identified burden and unmet needs in patients with cluster headache: An evidence-based qualitative literature review," *Cephalalgia Reports*, vol. 5, p. 25158163221096864, 2022.

[12] R. R. Buxbaum, "Everything won't be okay: The impact of therapists' justification of the mental healthcare system on their racially minoritized patients," Long Island University, Brooklyn, 2023.

[13] C. Guo, H. Ashrafian, S. Ghafur, G. Fontana, C. Gardner, and M. Prime, "Challenges for the evaluation of digital health solutions—A call for innovative evidence generation approaches," *NPJ Digit. Med.*, vol. 3, no. 1, p. 110, 2020.

[14] L. E. Lwakatare, A. Raj, I. Crnkovic, J. Bosch, and H. H. Olsson, "Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions," *Inf. Softw. Technol.*, vol. 127, p. 106368, 2020.

[15] K. D. Davis *et al.*, "Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities," *Nat. Rev. Neurol.*, vol. 16, no. 7, pp. 381–400, 2020.

[16] R. Gupta, S. Kumari, A. Senapati, R. K. Ambasta, and P. Kumar, "New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease," *Ageing Res. Rev.*, p. 102013, 2023.

[17] M. Yagi, K. Yamanouchi, N. Fujita, H. Funao, and S. Ebata, "Revolutionizing Spinal Care: Current Applications and Future Directions of Artificial Intelligence and Machine Learning," *J. Clin. Med.*, vol. 12, no. 13, p. 4188, 2023.

[18] U. Ullah and B. Garcia-Zapirain, "Quantum Machine Learning Revolution in Healthcare: A Systematic Review of Emerging Perspectives and Applications," *IEEE Access*, 2024.

[19] M. A. Myszczynska *et al.*, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nat. Rev. Neurol.*, vol. 16, no. 8, pp. 440–456, 2020.

[20] R. Thirunavukarasu, R. Gnanasambandan, M. Gopikrishnan, V. Palanisamy, and others, "Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review," *Comput. Biol. Med.*, p. 106020, 2022.

[21] E. Badidi, "Edge AI for early detection of chronic diseases and the spread of infectious diseases: opportunities, challenges, and future directions," *Futur. Internet*, vol. 15, no. 11, p. 370, 2023.

[22] N. K. Iyortsuun, S.-H. Kim, M. Jhon, H.-J. Yang, and S. Pant, "A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis," in *Healthcare*, 2023, vol. 11, no. 3, p. 285.

[23] J. Chung and J. Teo, "Mental health prediction using machine learning: taxonomy, applications, and challenges," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, pp. 1–19, 2022.

[24] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches," *Sensors*, vol. 24, no. 2, p. 348, 2024.

[25] J. R. A. Solares *et al.*, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures," *J. Biomed. Inform.*, vol. 101, p. 103337, 2020.

[26] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf. Sci. (Ny).*, vol. 513, pp. 429–441, 2020.

[27] H. El Haji *et al.*, "Evolution of Breast Cancer Recurrence Risk Prediction: A Systematic Review of Statistical and Machine Learning--Based Models," *JCO Clin. Cancer Informatics*, vol. 7, p. e2300049, 2023.

[28] W. H. Organization and others, "World mental health report: transforming mental health for all," 2022.

[29] D. Godinić and B. Obrenovic, "Effects of economic uncertainty on mental health in the COVID-19 pandemic context: social identity disturbance, job uncertainty and psychological well-being model," 2020.
[30] I. Coppola, N. Rania, R. Parisi, and F. Lagomarsino, "Spiritual well-being and mental health during the COVID-19 pandemic in Italy," *Front. Psychiatry*, vol. 12, p. 626944, 2021.
[31] S. Vatansever *et al.*, "Artificial intelligence and machine learning-aided drug discovery in central nervous system diseases: State-of-the-arts and future directions," *Med. Res. Rev.*, vol. 41, no. 3, pp. 1427–1473, 2021.
[32] S. Neethirajan, "Artificial Intelligence and Sensor Innovations: Enhancing Livestock Welfare with a Human-Centric Approach," *Human-Centric Intell. Syst.*, pp. 1–16, 2023.
[33] F. Sabah, Y. Chen, Z. Yang, M. Azam, N. Ahmad, and R. Sarwar, "Model optimization techniques in personalized federated learning: A survey," *Expert Syst. Appl.*, p. 122874, 2023.
[34] V. Garc\'\ia, J. S. Sánchez, A. I. Marqués, R. Florencia, and G. Rivera, "Understanding the apparent superiority of over-sampling through an analysis of local information for class-imbalanced data," *Expert Syst. Appl.*, vol. 158, p. 113026, 2020.
[35] M. E. Paoletti, O. Mogollon, S. Moreno, J. C. Sancho, and J. M. Haut, "A Comprehensive Survey of Imbalance Correction Techniques for Hyperspectral Data Classification," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 2023.
[36] Programmerrdai, "Mental Disorder Classification." 2021.