informatics

MDPI

Article

Explainable Artificial Intelligence for Workplace Mental Health Prediction

Tsholofelo Mokheleli *, Tebogo Bokaba and Elliot Mbunge

Department of Applied Information Systems, University of Johannesburg, Johannesburg 2006, South Africa; tbokaba@uj.ac.za (T.B.); elliotm@uj.ac.za (E.M.)
* Correspondence: tsholofelom@uj.ac.za

Abstract

The increased prevalence of mental health issues in the workplace affects employees' well-being and organisational success, necessitating proactive interventions such as employee assistance programmes, stress management workshops, and tailored wellness initiatives. Artificial intelligence (AI) techniques are transforming mental health risk prediction using behavioural, environmental, and workplace data. However, the "black-box" nature of many AI models hinders trust, transparency, and adoption in sensitive domains such as mental health. This study used the Open Sourcing Mental Illness (OSMI) secondary dataset (2016–2023) and applied four ML classifiers, Random Forest (RF), xGBoost, Support Vector Machine (SVM), and AdaBoost, to predict workplace mental health outcomes. Explainable AI (XAI) techniques, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), were integrated to provide both global (SHAP) and instance-level (LIME) interpretability. The Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalance. The results show that xGBoost and RF achieved the highest cross-validation accuracy (94%), with xGBoost performing best overall (accuracy = 91%, ROC AUC = 90%), followed by RF (accuracy = 91%). SHAP revealed that *sought_treatment*, *past_mh_disorder*, and *current_mh_disorder* had the most significant positive impact on predictions, while LIME provided case-level explanations to support individualised interpretation. These findings show the importance of explainable ML models in informing timely, targeted interventions, such as improving access to mental health resources, promoting stigma-free workplaces, and supporting treatment-seeking behaviour, while ensuring the ethical and transparent integration of AI into workplace mental health management.

Keywords: explainable artificial intelligence; machine learning; mental health; prediction; workplace

check for updates

Academic Editor: Olga Kurasova

Received: 16 August 2025
Revised: 4 October 2025
Accepted: 11 October 2025
Published: 26 November 2025

Citation: Mokheleli, T.; Bokaba, T.; Mbunge, E. Explainable Artificial Intelligence for Workplace Mental Health Prediction. *Informatics* **2025**, *12*, 130. https://doi.org/10.3390/informatics12040130

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The emotional, psychological, and social well-being of employees in the work environment is important [1] as it substantially affects how individuals cope with work demands, interact with colleagues, manage stress, and perform their roles effectively. Nevertheless, mental health problems are prevalent in workplaces despite the implementation of various workplace mental health interventions [2]. The World Health Organization (WHO) [3] defines mental health as "*a state of mental well-being that allows the person to cope with the stresses of life, realise his or her abilities, learn well, work well, and contribute to the community.*" Mental

health significantly impacts employee productivity, engagement, and retention in contemporary workplaces. For instance, poor mental health at work is associated with greater absenteeism, lower engagement, and lower productivity [4]. According to the WHO [3], mental health disorders are among the leading causes of disability in the workplace, resulting in the global economy losing approximately $1 trillion annually [5,6]. However, a friendly and enabling working environment with effective mental health and wellness initiatives promotes the emotional, psychological, and social well-being of employees and has relatively low staff turnover. Supportive work environments that support employee satisfaction generally have higher organisational performance levels [7]. For instance, organisations that invest in appropriate programmes for mental health report benefits such as higher morale among employees, more job satisfaction, and, most importantly, increased economic returns associated with investments through productivity gains and reductions in health costs [8,9]. The factors that result in poor mental health status in the workplace are complex and intertwined. Such factors include stress, high workloads, poor psychosocial working conditions, job insecurity, power imbalances and dynamics, poor practices by management, and limited support mechanisms [9–11]. In addition, bullying, job stressors, and discrimination in the workplace [12] contribute significantly to triggering anxiety, stress, burnout, distress, and depression [11].

However, while traditional mental health and wellness initiatives, such as employee assistance programmes, stress management workshops, and awareness campaigns, have attempted to address these challenges [2,3], they often lack predictive capabilities. Therefore, artificial intelligence (AI) techniques, especially machine learning (ML) algorithms, continue to transform how organisations monitor and respond to mental health risks. AI has brought a new and transformative capability into mental health by analysing large, complex datasets to extract patterns [13,14]. ML is a subfield of AI that focuses on developing algorithms and models that can learn patterns from data without being explicitly programmed. ML algorithms have been applied to solve different tasks, including regression, classification, and clustering problems [15]. ML models have been used for predicting mental health using data from workplace surveys, general practice, electronic health record systems, and social networks [16,17]. For instance, ML classification models such as logistic regression, decision trees, and random forest have also been used to analyse stress levels, job satisfaction, and risk factors in the workplace [17,18]. However, the "black-box" nature of traditional ML models applied to predicting mental health hinders trust, transparency, and their adoption in real practice. The lack of interpretability and explicit explanations of predicted outcomes remains a major barrier to ML adoption in healthcare settings [19]. Furthermore, many studies have used general mental health datasets without tailoring models to specific workplace contexts, leaving a gap in understanding the unique stressors and predictors in specific work environments [14,20]. Therefore, this study makes a significant contribution to the existing literature by addressing these gaps through the following:

- Developing and evaluating ML models using the Open Sourcing Mental Illness (OSMI) secondary dataset to predict workplace mental health outcomes.
- Applying Explainable AI (XAI) techniques (SHAP and LIME) to enhance model transparency and interpretability.

The remainder of the paper is organised as follows: Section 2 presents the related work on the application of ML in predicting mental health and the importance of XAI in mental health ML models. Section 3 describes the various methods used in this study. This includes the description of the data, preprocessing steps, ML models, XAI techniques, and performance evaluation metrics. Section 4 presents the results, while Section 5 discusses the results and findings. Implications of the study, as well as study limitations and future work, are presented in Section 5. Section 6 presents the conclusion.

2. Related Work

The intersection of mental health, ML, and XAI has gained significant attention due to the growing recognition of the impact of mental health on personal well-being and organisational performance. Although ML models have shown promising results in predicting mental health outcomes, their applications often lack transparency in the workplace and healthcare settings. This is because of the black box nature of AI models, which hinders trust, transparency, adoption, and usability in real health settings [21]. This section examines previous research on mental health in workplace environments, the use of ML in predicting mental health, and further analyses the importance of explainability in AI models.

The importance of employee mental health is increasingly recognised, with studies highlighting its influence on productivity, job satisfaction, and employee retention. Rautaray et al. [22] emphasised that creating a supportive workplace environment is important for enhancing employee well-being, reducing stigma, and improving organisational performance. Their study applied various ML models such as logistic regression (LR), k-nearest neighbours (KNN), decision trees (DT), and random forest (RF) classifiers to analyse employees' health data. The results of their study indicated that the DT classifier was the most effective, achieving 81% accuracy. Their study also highlighted that work–life balance, social support, and job satisfaction influence mental health and suggested further exploration of additional variables such as physical health and lifestyle habits. Similarly, Lohia et al. [13] used the OSMI Mental Health in Tech Survey dataset to investigate workplace conditions affecting employees' mental and physical well-being. Their findings highlighted the potential of encouraging workplace camaraderie and implementing health campaigns to reduce mental stress. However, their study did not explore the scalability and adaptability of their models to diverse workplace settings. Almaleh [17] expanded on the workplace mental health literature by predicting whether employees would seek mental health treatment using RF, LR, support vector machines (SVM), AdaBoost, and gradient boosting (xGBoost). The study used the OSMI Mental Health in Tech Survey dataset and recorded high accuracy values, ranging from 82.83% to 85.23%. The boosting classifiers, AdaBoost and xGBoost, had the highest accuracy at 85.03% and 85.23%, respectively. Almaleh [17] further revealed that over half of the employees surveyed faced psychological issues, and the study emphasised the importance of flexible scheduling and tailored wellness programmes.

Despite these contributions, the integration of explainability in ML models for workplace mental health remains underexplored. Chauhan and Renjith [18] introduced a comprehensive stress detection and mitigation system that uses convolutional neural networks (CNN), KNN, and image processing techniques for real-time monitoring of workplace stress. Their approach integrates data collection and user interfaces to provide personalised stress management solutions, aiming to promote healthier work environments. However, their work emphasised the need to address ethical considerations, such as privacy and fairness, highlighting the importance of incorporating XAI to ensure responsible implementation and foster user trust.

The application of ML in mental health has explored diverse datasets and prediction tasks. Graham et al. [14] conducted a review of 28 studies that used electronic health records, mood rating scales, and social media data to predict mental health outcomes, including depression and suicidal ideation. The review indicated that some models predicted mental health outcomes with high accuracy. However, limitations such as data quality and sample size consequently affected the performance of some ML models, which frequently led to overfitting. Furthermore, Graham et al. [14] highlighted the importance of balanced datasets and algorithmic transparency, especially when addressing rare events such as

suicidal ideation. Srividya et al. [23] explored clustering and classification techniques for identifying mentally distressed individuals by employing SVM, DTs, LR, Naïve Bayes, and ensemble methods. They used 656 samples collected from two populations (300 from population 1 and 356 from population 2), covering high school students, college students, and working professionals. Class labels were derived through clustering and validated using the Quality of Experience approach through Mean Opinion Score, with LIME applied for trust computation. Among the models tested, ensemble methods, specifically Bagging and Tree Ensemble (RF), achieved the highest accuracy at 90%, highlighting the effectiveness of these approaches and the value of feature subset selection in improving classifier reliability. Similarly, Mallick and Panda [24] applied feature selection algorithms such as Recursive Feature Elimination (RFE) and Recursive Feature Elimination with Cross-Validation (RFECV), and Least Absolute Shrinkage and Selection Operator (LASSO) to enhance the predictive power of ML models on the OSMI dataset. Among the classifiers tested, RF frequently outperformed others, achieving an accuracy of 78%, with RFECV combined with RF reaching 79%, indicating that technical workers were more likely to experience mental health issues than non-technical workers. While their study demonstrated the benefits of feature selection in improving model performance, it did not explicitly discuss the study's limitations, leaving room for further exploration of potential constraints such as dataset representativeness and generalisability. Nanath et al. [20] highlighted the significant impact of COVID-19 on mental health, which led to increased levels of depression, loneliness, and stress across the United States, Germany, and the United Kingdom. Their study introduced a global mental health index using social media data to enable scalable and cost-effective trend analysis compared to traditional methods. Nanath et al. [20] suggested that physical work environments reduce depressive symptoms by fostering belonging and professional confidence. However, their study faced limitations such as reliance on Twitter data and potential biases in mobility metrics. Sujal et al. [25] applied multiple ML models, including KNN, LightGBM, and xGBoost, to explore workplace stress and its determinants using the OSMI Mental Health in Tech Survey dataset. Among these, KNN achieved the highest accuracy (84.23%), followed closely by LightGBM (84.22%) and xGBoost (84.95%). Their findings linked urban work environments to higher mental health challenges. Kapoor and Goel [26] also addressed workplace stress using ML, including LR, KNN, DT, RF, xGBoost, SVM, and the OSMI Mental Health Survey data (2014 and 2021) to identify stress management needs among professionals. Data preprocessing involved cleaning and reducing the dataset to 19 key attributes, focusing on personal and professional factors. DT performed best on the 2021 dataset with an accuracy of 82%, and AdaBoost excelled on the 2014 dataset with an accuracy of 81%.

While existing models prioritise predictive accuracy, they often have limited focus on the explainability of models, which limits their application in real health settings, where interpretability is essential for stakeholder trust. The explainability of AI models is essential in mental health predictions to enhance trust and facilitate clinical decision-making. Chung and Teo [27] identified the absence of explainable models as a key limitation in current ML applications for mental health. Explainable AI seeks to address the interpretability and transparency gaps of ML models by making them interpretable [28]. XAI offers a pathway to build the trust and transparency required in the pursuit of effective responses to workplace mental health challenges [29,30]. XAI techniques such as LIME and SHAP help to understand features in the model that most drive the predictions. For instance, SHAP assigns an importance score to the individual predictors [31]. This helps identify key job stressors, such as workload or lack of social support, that contribute to mental health problems. For example, Srividya et al. [23] applied LIME to improve classifier transparency.

XAI can bridge the gap between predictive power and interpretability [29]. However, the broader adoption of XAI remains limited.

Table 1 provides an overview of studies that applied ML techniques to predict mental health. The *References* column indicates the studies by author; the *AI/ML Techniques* column highlights the algorithms used or, in the case of review articles, the most applied methods; the *Performance Metrics* column outlines the evaluation criteria; the *Data/Data Description* column describes the datasets used or provides additional details when the name is unspecified; the *Focus Area* column specifies the context or objective of the study; and finally, the *XAI Applied* column indicates whether explainable AI methods were applied. The table reveals a notable scarcity of explainable AI models in the reviewed literature. Another recurring pattern is the frequent use of the OSMI Mental Health in Tech Survey dataset alongside popular ML techniques such as DTs, RFs, and KNN. This trend highlights the pressing need for greater integration of explainability to enhance the trustworthiness and practical application of the models in real-world mental health scenarios.

Table 1. Summary of related studies on ML applications in mental health.

Studies	AI/ML Techniques	Performance Metrics	Data/Data Description	Focus Area	XAI Applied
Rautaray et al. [22]	LR, KNN, DT, RF	Accuracy, Classification error, False Positive Rate, Precision, Confusion Matrix	It is not mentioned, but comprises employee behaviour, health, and work–life impact.	Employee mental health and well-being	Not applied
Lohia et al. [13]	KNN, LR, DT, RF, Bagging, Boosting	Precision, Recall, Classification Accuracy, F1 Score, AUC	OSMI Mental Health in Tech Survey	Workplace mental health and operational efficiency	Not applied
Graham et al. [14]	Supervised, unsupervised ML, Deep Learning, and Natural Language Processing	AUC, Classification Accuracy, Sensitivity, Specificity, Precision, F1 Score	EHRs, Mood Scales, Brain Imaging, social media 28 research articles reviewed	General mental health predictions	-
Srividya et al. [23]	Clustering (K-means), SVM, DT, NB, KNN, LR	Confusion Matrix, Classification Accuracy, Precision, Recall, F-beta Score	656 samples, 20 features, 3 class labels	Identifying mentally distressed individuals	Used LIME for trust computation
Nanath et al. [20]	Lasso, RFE, RFECV, Naïve Bayes, SGD, KNN, DT, RF, SVM, NN, XGB	Accuracy, Recall, Precision	Twitter data, Google mobility data, Lockdown index data	COVID-19 mental health impacts	Not applied
Almaleh [17]	RF, SVM, LR, AdaBoost, GB	Accuracy, Precision, Recall, Confusion Matrix, AUC, ROC Curve	OSMI dataset	Mental health treatment prediction	Not applied
Chung et al. [27]	-	Accuracy, F-measure, AUC	30 research articles reviewed	Mental health prediction techniques	Not applied
Sujal et al. [25]	LR, KNN, DT, RF, NB, SVM, AdaBoost, GB, Light GBM	Precision, Recall, F1 Score, Accuracy	OSMI Mental Health Survey	Workplace stress and mental health	Not applied
Kapoor et al. [26]	LR, KNN, DT, RF, GB, AdaBoost, XGB, SVM	Classification Accuracy, False Positive Rate, Precision, AUC, ROC Curve	OSMI Mental Health Survey (2021 and 2014)	Stress management in tech professionals	Not applied
Mallick et al. [24]	LR, KNN, DT, RF, NN, XGB	Accuracy, Recall, Precision	Tech Survey dataset	Mental health feature selection	Not applied
[18]	CNN, KNN	Accuracy, Precision, Recall	Stress detection system data	Real-time stress detection and management	Not applied

Although existing literature has made notable progress in using ML to predict mental health outcomes, the lack of explainability in these models presents significant challenges to their practical application. A survey by Mienye et al. [32] and a review study by Noor et al. [33] showed that XAI has become important in healthcare, enabling interpretability and trust in ML models. Nevertheless, applying these techniques to domain-specific contexts such as mental health remains challenging. Recent studies have begun to address this gap; for instance, Atlam et al. [34] proposed an explainable ensemble framework for predicting mental health disorders, while Ahmed et al. [35] demonstrated how XAI methods

can improve the interpretability of depression detection and severity assessment. Despite these advances, many studies still emphasise accuracy while overlooking transparency, ethical considerations, and user-centred evaluation [13,22,24]. Furthermore, different job stressors associated with workplace environments demand customised models that incorporate domain-specific factors. This paper seeks to bridge these gaps by integrating XAI techniques with workplace-focused datasets, providing interpretable models for predicting mental health risks while balancing accuracy and transparency.

3. Methods

3.1. Dataset Overview and Data Preprocessing

The secondary dataset used in this study was sourced from the OSMI Mental Health in Tech Survey. The dataset is publicly available and has been cited in various research studies [36]. The novelty of this research lies in its incorporation of data, spanning from 2016 to 2023. Although the dataset spans both pre-COVID and post-COVID years, the study retained the full 2016 to 2023 period to capture long-term variations in workplace mental health. This approach ensured that the models generalise across diverse conditions and reflect the broader evolution of workplace mental health risks, including but not limited to pandemic-related effects. To the best of our knowledge, this is the first study to use data for this extended time frame. After consolidating the annual survey data, the final dataset included 3439 instances and 47 features, including the target class (as shown in Table 2). Although the number of survey questions varied each year, the 47 features selected for this study were chosen based on their relevance to the research objective, and they were common throughout all eight surveys, with some minor wording differences. See Table 3 for the original questions in the dataset and the simplified terms for the analysis.

Table 2. Overview of the dataset instances and columns by year.

Year	Number of Instances	Number of Columns
2016	1433	63
2017	756	123
2018	417	123
2019	352	82
2020	180	120
2021	131	124
2022	164	126
2023	6	126

Table 3. Survey questions and feature names.

Survey Questions	Feature Name
Are you self-employed?	Self_employed
How many employees does your company or organization have?	Company_size
Is your employer primarily a tech company/organization?	Employer_tech
Is your primary role within your company related to tech/IT?	Role_tech
Does your employer provide mental health benefits as part of healthcare coverage?	Mh_benefits
Do you know the options for mental health care available under your employer-provided health coverage? *Do you know the options for mental health care available under your employer-provided coverage?*	Know_mh_options

Table 3. *Cont.*

Survey Questions	Feature Name
Has your employer ever formally discussed mental health (for example, as part of a wellness campaign or other official communication)?	Employer_discussed_mh
Does your employer offer resources to learn more about mental health disorders and options for seeking help? *Does your employer offer resources to learn more about mental health concerns and options for seeking help?*	Mh_resources
Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources provided by your employer?	Anonymity_protected
If a mental health issue prompted you to request a medical leave from work, how easy or difficult would it be to ask for that leave? *If a mental health issue prompted you to request a medical leave from work, asking for that leave would be:*	Ease_medical_leave
Would you feel comfortable discussing a mental health issue with your direct supervisor(s)?	Comfort_supervisor
Would you feel comfortable discussing a mental health issue with your coworkers?	Comfort_coworkers
Overall, how much importance does your employer place on mental health? *Do you feel that your employer takes mental health as seriously as physical health?*	Employer_importance_mental
Do you know local or online resources to seek help for a mental health issue? *Do you know local or online resources to seek help for a mental health disorder?*	Know_resources_mh
If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to clients or business contacts?	Reveal_to_clients
If you have revealed a mental health disorder to a client or business contact, how has this affected you or the relationship? *If you have revealed a mental health issue to a client or business contact, do you believe this has impacted you negatively?*	Client_relationship_impact
If you have been diagnosed or treated for a mental health disorder, do you ever reveal this to coworkers or employees?	Reveal_to_coworkers
If you have revealed a mental health disorder to a coworker or employee, how has this impacted you or the relationship? *If you have revealed a mental health issue to a coworker or employee, do you believe this has impacted you negatively?*	Coworker_relationship_impact
Do you believe your productivity is ever affected by a mental health issue?	Productivity_affected
If yes, what percentage of your work time (time performing primary or secondary job functions) is affected by a mental health issue?	Productivity_percentage
Do you have previous employers?	Prev_employers
Have your previous employers provided mental health benefits?	Prev_employer_benefits
Were you aware of the options for mental health care provided by your previous employers?	Prev_employer_options
Did your previous employers ever formally discuss mental health (as part of a wellness campaign or other official communication)?	Prev_employer_discussed_mh
Did your previous employers provide resources to learn more about mental health disorders and how to seek help?	Prev_employer_resources
Was your anonymity protected if you chose to take advantage of mental health or substance abuse treatment resources with previous employers?	Prev_anonymity_protected
Would you have been willing to discuss your mental health with your direct supervisor(s)? *Would you have been willing to discuss a mental health issue with your direct supervisor(s)?*	Prev_comfort_supervisor
Would you have been willing to discuss your mental health with your coworkers at previous employers? *Would you have been willing to discuss a mental health issue with your previous co-workers?*	Prev_comfort_coworkers

Table 3. *Cont.*

Survey Questions	Feature Name
Overall, how much importance did your previous employer place on mental health? *Did you feel that your previous employers took mental health as seriously as physical health?*	Prev_employer_importance_mental
Do you currently have a mental health disorder?	Current_mh_disorder
Have you ever been diagnosed with a mental health disorder? *Have you been diagnosed with a mental health condition by a medical professional?*	Diagnosed_mh *
Have you had a mental health disorder in the past?	Past_mh_disorder
Have you ever sought treatment for a mental health disorder from a mental health professional? *Have you ever sought treatment for a mental health issue from a mental health professional?*	Sought_treatment
Do you have a family history of mental illness?	Family_history
If you have a mental health disorder, how often do you feel that it interferes with your work when being treated effectively? *If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?*	Interference_work_effective_treatment
If you have a mental health disorder, how often do you feel that it interferes with your work when NOT being treated effectively (i.e., when you are experiencing symptoms)? *If you have a mental health issue, do you feel that it interferes with your work when NOT being treated effectively?*	Interference_work_no_treatment
Have your observations of how another individual who discussed a mental health issue made you less likely to reveal a mental health issue yourself in your current workplace? *Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?*	Observation_discouraged_revealing
How willing would you be to share with friends and family that you have a mental illness?	Willing_share_with_friends_family
Would you be willing to bring up a physical health issue with a potential employer in an interview?	Willing_physical_issue_interview
Would you bring up your mental health with a potential employer in an interview? *Would you bring up a mental health issue with a potential employer in an interview?*	Willing_mh_interview
Has being identified as a person with a mental health issue affected your career? *Do you feel that being identified as a person with a mental health issue would hurt your career?*	Impact_on_career
If they knew you suffered from a mental health disorder, how do you think that your team members/co-workers would react? *Do you think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue?*	Coworker_reaction
Have you observed or experienced an unsupportive or badly handled response to a mental health issue in your current or previous workplace?	Observed_unsupportive_response
What is your age?	Age
What is your gender?	Gender
What country do you live in?	Country_of_residence
What country do you work in?	Country_of_work

* Target class.

The target variable, *Diagnosed_mh*, was chosen because it aligns with the study's objective Although *Past_mh_disorder* and *Current_mh_disorder* are correlated, they were retained to provide contextual information about prior or ongoing mental health conditions, as they are known predictors of workplace psychological distress, absenteeism, and performance declines [37]. SHAP and LIME were applied to assess and interpret the variables contributions.

The original survey questions were converted into concise and easily understandable feature names to improve readability and interpretability. The wording of the surveys from 2017 to 2023 remained constant, but 2016 used different wording. The 2016 questions are italicised in Table 3.

Data preprocessing steps were implemented to ensure the dataset was prepared for further analysis and modelling. This process addressed inconsistencies and optimised the dataset for analysis, with key steps including handling missing data, label encoding, and data cleaning. These steps are explained in the subsequent section.

Missing data was addressed using Multiple Imputation by Chained Equations (MICE), a method proven to be highly effective for this dataset in prior studies [5]. MICE iteratively imputes missing values by predicting them based on other observed features, preserving the overall integrity of the dataset [38]. Survey responses were converted into numeric formats using label encoding. Categorical variables were label-encoded to enable model compatibility. Since these categories have no inherent order, encoding did not imply ordinal meaning. In addition, data cleaning was applied to features such as age and gender by removing outlier values and standardising gender responses into three categories ("Male", "Female", and "Other") to ensure consistency and analytical relevance. Responses outside the range of 15 to 70 years were excluded for age. The lower limit age of 15 aligns with the legal working age as stated in the International Labour Organization (ILO) minimum age convention of 1973 (No.138) [39]. The cutoff at 70 ensures relevance to workplace contexts, considering the average retirement age of 65, while including those active up to retirement, minimising outlier effects. Data imputation was done on all characteristics except the target class during data preparation to ensure the target variable's integrity. Instances without target values were dropped, resulting in a dataset of 2240 instances accessible for model training and evaluation, down from 3439 initially. The dataset was then divided into training and testing sets, with 80% used for training and 20% for testing, using stratified sampling to preserve class distribution across both subsets [40]. This split guaranteed that the models were trained on a large percentage of the data while still leaving enough instances for an unbiased performance evaluation.

This study used an anonymized secondary OSMI dataset, which is publicly available and accessible. No direct contact with participants was involved; therefore, no consent forms were used to access the public domain secondary OSMI Mental Health data. The secondary data can be accessed in this repository (website): Open Sourcing Mental Illness (OSMI) Research, URL: https://osmihelp.org/research.html (accessed on 4 July 2025). In addition, this study was granted ethical approval by the Department of Applied Information Systems Research Ethics Committee (AIS REC) at the University of Johannesburg (ethical clearance code: 2023AIS009).

3.2. Exploratory Data Analysis

Exploratory data analysis (EDA) was used to understand the dataset and its distributions. It is an important phase in research as it examines data for distributions, outliers, and anomalies, which helps guide hypothesis testing. This technique frequently involves visualising and analysing data through graphical representations [41]. Univariate analyses were performed on key features such as age, gender, and class distribution. Figure 1 shows the age distribution, revealing that most respondents were between 25 and 40 years old. This age group represents the largest segment of the working-age population, offering valuable insights into workplace mental health trends. Gender distribution is shown in Figure 2, with 2369 male respondents, 937 female respondents, and 99 individuals identifying as "Other." This distribution reflects the male predominance typical of the tech industry. Figure 3 shows the class distribution within the dataset, highlighting a notable imbalance.

Approximately 66.9% of respondents were classified as "Yes" for the target variable, while 33.1% were classified as "No." This imbalance emphasises the need for suitable sampling techniques to handle imbalanced data during model training to ensure fair and effective performance across all classes.

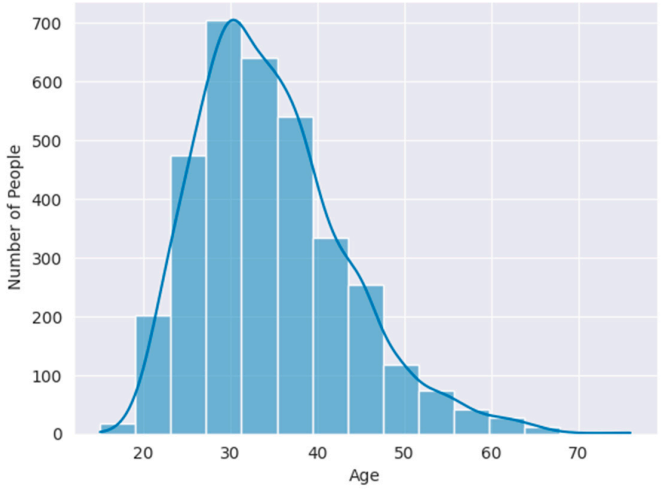


Figure 1. Age distribution.

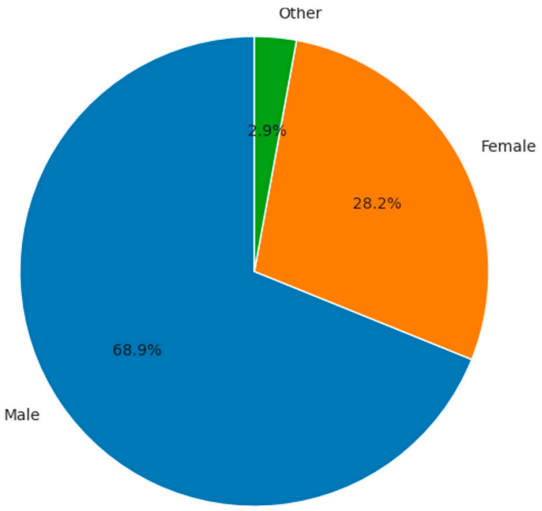


Figure 2. Gender distribution.

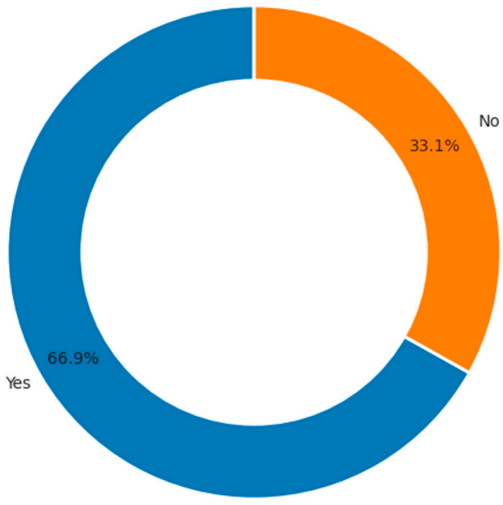


Figure 3. Class distribution.

Figure 4 presents a geographic analysis, which maps the distribution of respondents worldwide. The top countries in the dataset include the United States with 2059 respondents, the United Kingdom with 325, and Canada, Germany, and India with 147, 110, and 98 respondents, respectively. This analysis highlights the global nature of the dataset, with a substantial proportion of respondents from the United States and other English-speaking countries.

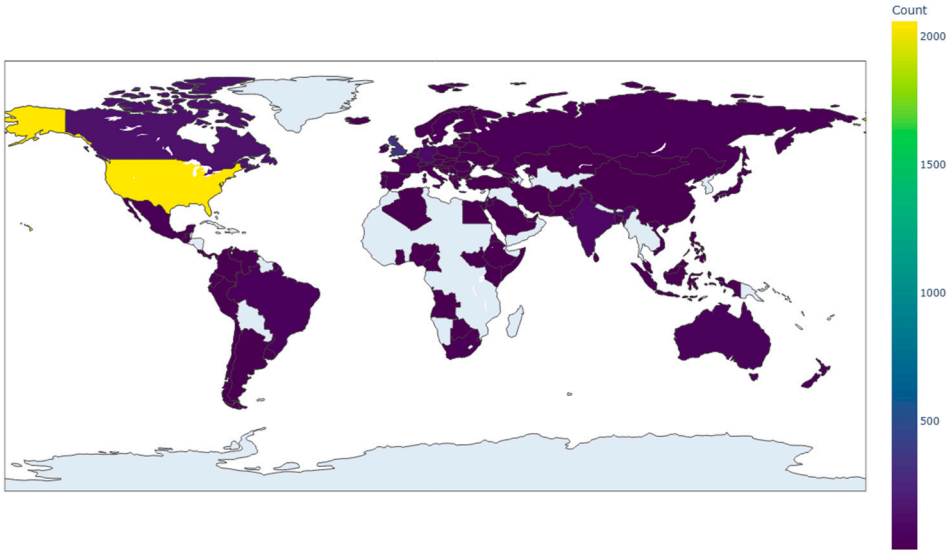


Figure 4. Global distribution of respondents based on workplace location.

3.3. Class Imbalance

EDA identified a large class imbalance in the target variable *Diagnosed_mh*. Initially, there were 588 instances representing Class 0 (those who had not been diagnosed with a mental health disorder) and 1204 instances representing Class 1 (those who were diagnosed with a mental health disorder). This imbalance may skew the performance of the model by favouring the majority class. To address this issue, the synthetic minority over-sampling technique (SMOTE) was used. SMOTE is a widely validated oversampling method that balances minority and majority classes by generating synthetic samples rather than duplicating existing ones. This helps ML models better learn decision boundaries. SMOTE is robust and widely used to address data imbalance in various public health studies [16,42,43]. In mental health, various authors, including Alghamdi et al. [42] and Sowjanya and Mrudula [43], applied SMOTE to address class imbalance before implementing predictive models. SMOTE resulted in balanced classes of 1204 instances in both Class 0 and Class 1. This resampling ensured equitable representation in the training data.

3.4. Model Development

This section explores the ML classifiers, performance metrics, and XAI techniques used. It provides a rationale for their selection, emphasising the balance between predictive performance and interpretability to ensure actionable insights for workplace stakeholders.

3.4.1. ML Classifiers

This study applied RF, xGBoost, SVM, and adaptive boosting (AdaBoost) to predict mental health outcomes, which were implemented in Python 3.12, using scikit-learn (v1.6.1), xGBoost (v3.0.5), and NumPy (v2.0.2) for numerical computation. The *RF algorithm* was chosen for its great accuracy by combining several DTs. The RF ensemble structure makes it challenging to separate the impact of various features without using explainability tools [44]. xGBoost (XGB) was selected for its ability to handle non-linear connections robustly using sophisticated methods, including feature trimming and regularisation [45].

Although it has advantages, the multi-layered decision-making process of XGB requires applying XAI techniques to improve prediction transparency and confidence. *SVMs* can effectively separate overlapping classes by mapping data into high-dimensional spaces using kernels [44]. *AdaBoost* was considered because of its iterative strategy for enhancing accuracy by concentrating on data points that were incorrectly classified [46]. Although efficient, the complexity brought about by its dynamic weighting mechanism necessitates using explanation and visualisation tools to guarantee understandability. These models were selected after thoroughly analysing relevant literature and predictive power, and their capacity to handle the complex, non-linear interactions in workplace mental health data.

3.4.2. Performance Metrics

This study used various performance evaluation metrics such as accuracy, precision, recall, F1-score, ROC AUC, geometric mean, balanced accuracy, and a confusion matrix to assess the performance of mental health ML models (See Tables 4 and 5). *Accuracy* indicates how well a model performs by calculating the percentage of accurate predictions made in all cases [47]. *Recall* assesses the capacity of the model to identify every real positive case, and *precision* measures the percentage of accurately detected positive cases among all anticipated positives [48]. The *F1-score* is good for unbalanced datasets since it balances precision and recall by taking the harmonic mean of these two metrics [49]. Plotting the true positive rate against the false positive rate at different thresholds allows the *ROC AUC* to assess the ability of the model to discriminate across classes; higher values indicate better discriminatory ability. The *geometric mean* ensures that both positive and negative class predictions are given equal weight by providing a balanced perspective on sensitivity and specificity [49]. Compared to standard accuracy, *balanced accuracy* provides a more comprehensive assessment by averaging sensitivity and specificity to account for class imbalances [47].

Table 4. Performance metrics and formulas.

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F1 Score	$2 \times \frac{Precision * Recall}{Precision+Recall}$
AUC	$Integral \int_0^1 ROC$
Geometric mean	$\sqrt{Sensitivity * Specificity}$
Balanced Accuracy	$\frac{Sensitivity+Specificity}{2}$

Table 5. Binary classification confusion matrix.

	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

A comprehensive understanding of the model's prediction is made possible by using a 2×2 *confusion matrix*—a tabular representation of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) (as shown in Table 5). This matrix offers important information for additional model improvements, and it helps identify trends in misclassifications.

3.4.3. Explainable AI Used

The study incorporated XAI techniques such as SHAP and LIME to improve the transparency of the ML models. SHAP provides a consistent measure of feature relevance by providing a Shapley value to each feature to indicate its contribution to the overall prediction of the model [31]. This approach evaluates how specific variables influence predictions across multiple instances. LIME enhances SHAP by providing locally interpretable reasons for specific predictions. It operates by perturbing the input data and measuring the changes in model output, resulting in an interpretable surrogate model for each prediction [50]. This technique helps to understand the reasoning behind certain predictions.

3.4.4. Experimental Setup

The Google Colab platform was used for the experiment, taking advantage of its computational power for effective processing and model training. The study compared the performance of four ML models: RF, xGBoost, SVM, and AdaBoost. To solve class imbalance, SMOTE was used, and hyperparameter tuning was performed using GridSearchCV to improve model performance and minimise overfitting. Models were trained and tested based on performance metrics such as accuracy, precision, recall, and F1 score, and their robustness was validated using k-fold cross-validation. To improve interpretability, XAI techniques, SHAP, and LIME, were used to provide insights into feature importance and individual prediction explanations. This methodical experimental setting supports a thorough examination of the OSMI secondary dataset and ML models, as shown in Figure 5.

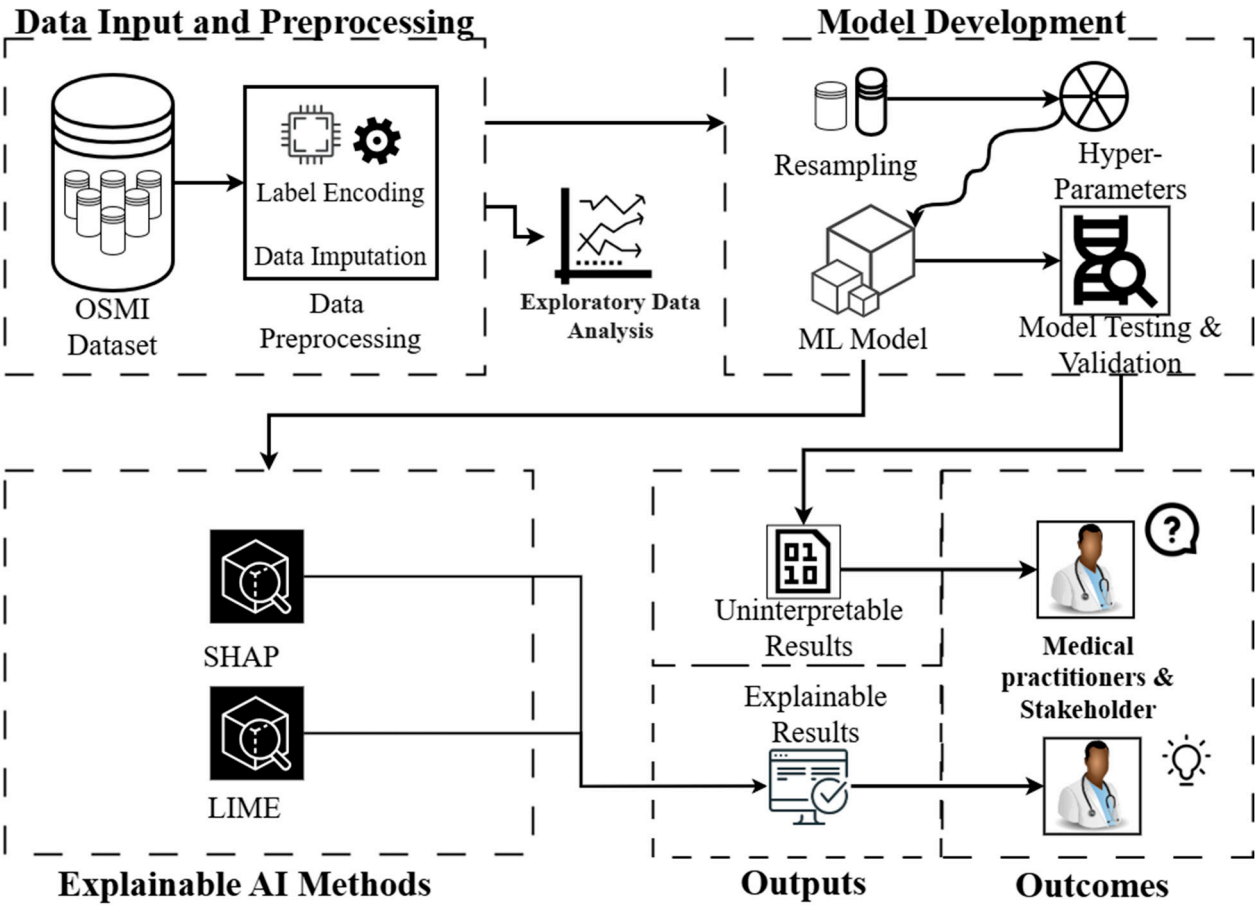


Figure 5. An overview of the analysis.

4. Results

This section presents the findings from the analysis of the OSMI dataset, highlighting key demographic characteristics and mental health patterns among participants. It further presents predictors of mental health at the workplace and subsequently summarises the models' performance evaluation metrics and explainability results.

4.1. Performance Results of the ML Models

The performance evaluation metrics were calculated based on the values of the confusion matrix. The confusion matrix shows the true positives, true negatives, false positives, and false negatives. This provides insights and trade-offs between sensitivity and specificity for each model. This is consistent with the goal of the study of creating interpretable ML models that reliably predict mental health outcomes and uncover actionable factors. The study applied GridSearchCV to optimise hyperparameters for each classifier using 5-fold cross-validation (Table 6). This systematic approach allowed for the identification of the ideal parameter settings for each model, improving their predictive performance. The models were then tested using a set of performance metrics, as shown in Figure 6, and a confusion matrix, as shown in Figures 7–10.

Table 6. Best hyperparameter settings for each classifier.

Model	Best Parameters
RF	n_estimators: 200; max_depth: 20; min_samples_split: 5; min_samples_leaf: 2; bootstrap: True
xGBoost	max_depth: 3; learning_rate: 0.01; n_estimators: 500; subsample: 1.0; colsample_bytree: 1.0
SVM	C: 0.1; kernel: linear; gamma: scale
AdaBoost	n_estimators: 100; learning_rate: 0.01

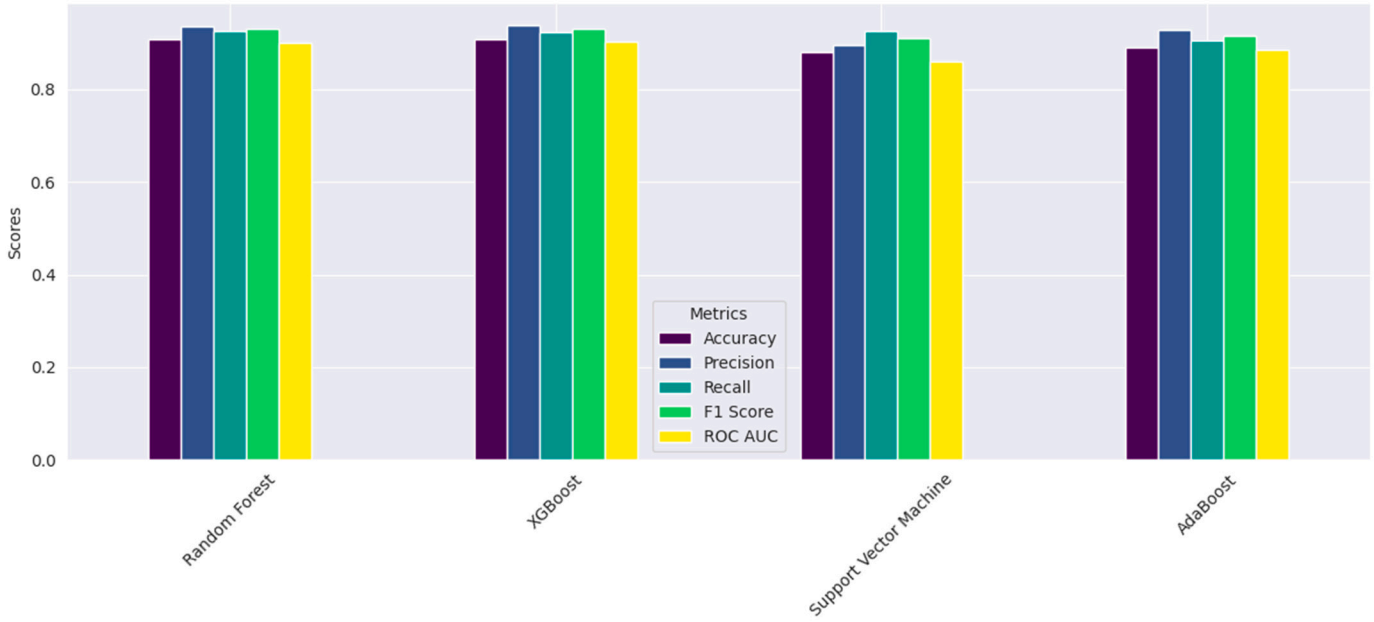


Figure 6. Results of the ML models.

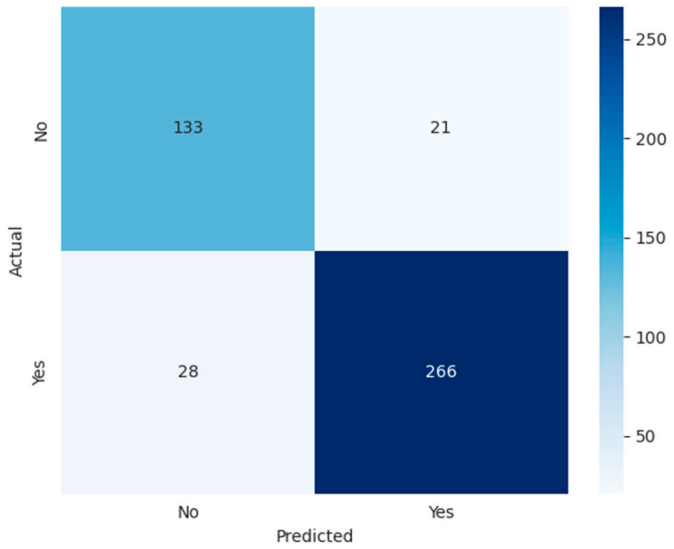


Figure 7. Confusion matrix for AdaBoost.

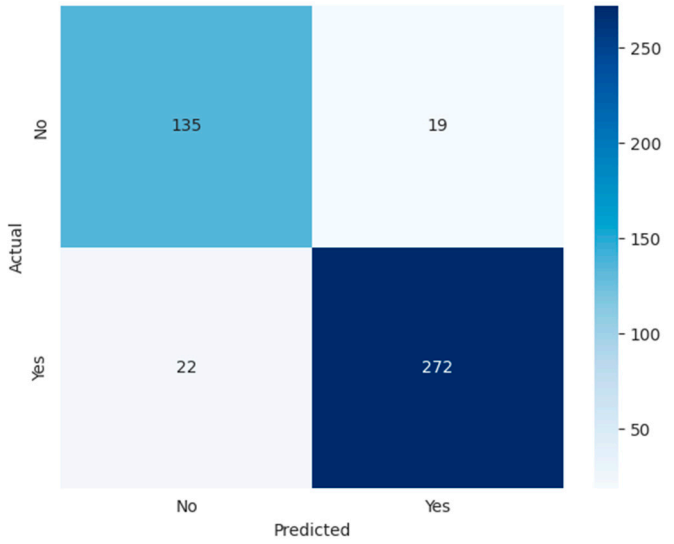


Figure 8. Confusion matrix for RF.

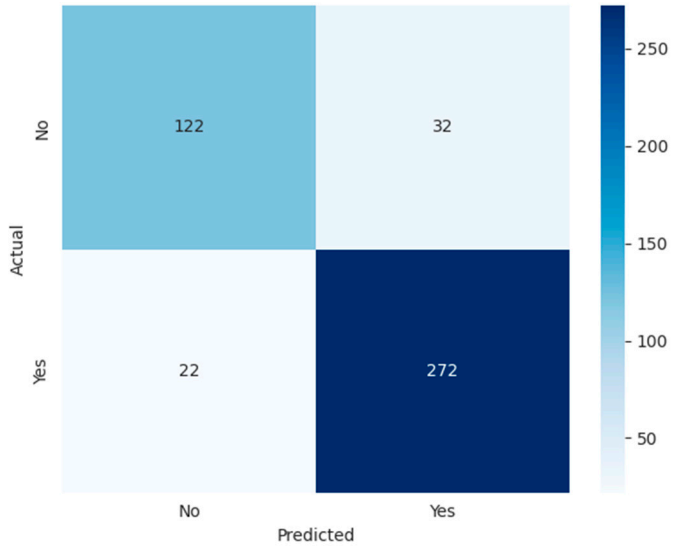


Figure 9. Confusion matrix for SVM.

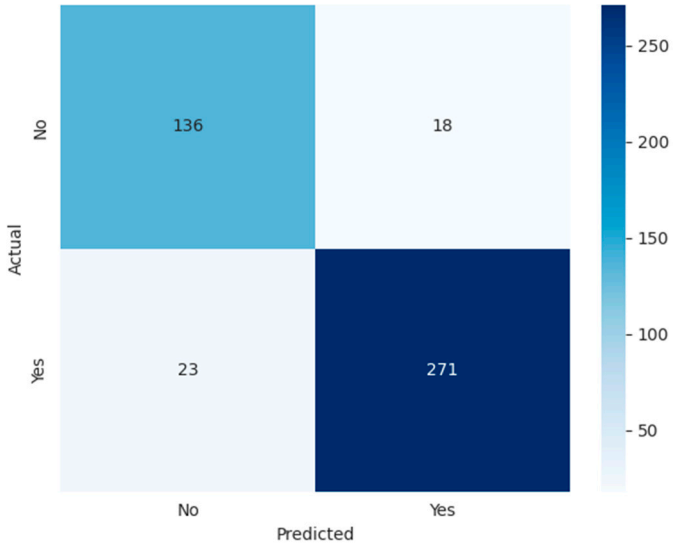


Figure 10. Confusion matrix for xGBoost.

From the evaluation metrics shown in Figure 6, xGBoost and RF performed the best overall, achieving high accuracy (91%) and balanced accuracy (90%). xGBoost achieved greater precision (94%) and the best ROC AUC (90%), indicating its ability to effectively differentiate between classes. In addition, its recall (92%) and F1 score (93%) indicate balanced performance across all criteria, placing it as a strong contender for the top model. RF performed similarly to xGBoost in most metrics, with slightly lower precision (93%) and ROC AUC (90%). Its high recall (93%) and geometric mean (90%) suggest that such xGBoost suits class imbalances well. SVM has the highest recall (93%), indicating its effectiveness in detecting true positives. However, its lower accuracy (88%), precision (89%), and ROC AUC (86%) indicate limits in its overall predictive balance. While constant across metrics, AdaBoost achieved significantly lower accuracy (89%) and F1 score (92%), making it less competitive than xGBoost and RF. Overall, xGBoost emerged as the best-performing model due to its balanced performance across all criteria, while RF follows closely. The decision between these models is determined by the unique application requirements, such as whether precision or recall should be prioritised.

The confusion matrices for the assessed models provide essential information about their effectiveness in classifying mental health outcomes. As shown in Figure 7, AdaBoost accurately identified 133 cases of Class 0 (no diagnosis) and 266 instances of Class 1 (diagnostic), with 21 false positives and 28 false negatives. This demonstrates an outstanding balance between sensitivity (recall for Class 1) and specificity (recall for Class 0), indicating reliability in detecting both groups. The RF model (Figure 8) fared similarly well, detecting 135 real Class 0 and 272 true Class 1 cases. It generated 19 false positives and 22 false negatives, slightly surpassing the overall sensitivity of AdaBoost. Its balanced performance in both classes demonstrates its suitability for this role. Figure 9 shows that the SVM has the maximum sensitivity, correctly categorising 272 Class 1 instances. However, its lower specificity results in a larger false positive count (32) and 122 accurate Class 0 predictions. This imbalance suggests that while SVM excels at recognising positive situations, it struggles to reduce false alarms. xGBoost (Figure 10) performs similarly to RF, properly recognising 136 Class 0 and 271 Class 1 occurrences. It reported 18 false positives and 23 false negatives, demonstrating its good predictive ability and balance of sensitivity and specificity.

Furthermore, the study validated the performance of the models using 5-fold cross-validation. This technique ensures that a complete dataset is used for both training and testing, which improves model robustness and generalisability [51]. Unlike the previously

described 80:20 split, cross-validation uses the entire dataset by iteratively training on 80% of it and testing on the remaining 20% in each fold. This method provides a more detailed assessment of the model's performance. The average cross-validation (CV) accuracy and standard deviation for each classifier are depicted in Table 7.

Table 7. Cross-validation results.

Classifier	Mean CV Accuracy	Standard Deviation
Random Forest	94%	0.01
xGBoost	94%	0.01
SVM	91%	0.02
AdaBoost	91%	0.01

Following cross-validation, RF and xGBoost remained the best-performing models, with a mean accuracy of 94% and a standard deviation of 0.01, showing consistent and robust performance across multiple folds. AdaBoost also managed well, with a mean accuracy of 91% and a standard deviation of 0.01. However, it fell short of the best models in accuracy. The SVM had a respectable mean accuracy of 91%, but the most outstanding standard deviation (0.02), indicating more variability and less steady performance over folds.

Overall, RF and xGBoost obtained the best accuracy and consistency, which makes them the most reliable classifiers in this study. The final choice depends on other considerations, such as interpretability and computational efficiency.

4.2. Explainable AI

This study applied two widely used explainable XAI techniques, SHAP and LIME, to provide explanations.

4.2.1. SHAP Analysis

SHAP is a widely used XAI method that helps interpret ML model predictions by showing how each feature contributes to a specific outcome. It assigns each feature an importance value for a particular prediction [52]. In this study, SHAP was applied to the RF and xGBoost models to ensure a balance between high performance and interpretability, thereby developing an interpretable ML model for workplace mental health predictions. The *TreeExplainer* was used to construct SHAP values based on the test data, allowing stakeholders to understand the underlying predictors and the reasoning behind the predictions. The summary, force, and bar plots were key visualisations that showed the most important workplace-related features impacting predictions. These visualisations provided actionable insights into the RF and xGBoost models' decision-making processes, enhancing their transparency and interpretability.

Since xGBoost was the best-performing model, the SHAP analysis presented in this study focuses on xGBoost, as it shares similar predictors with the RF, the second-best-performing model. SHAP values were used to interpret the model's predictions, revealing the role of each feature in predicting mental health outcomes. The SHAP summary plot in Figure 11 shows that key predictors such as *Sought_treatment*, *Past_mh_disorder*, and *Interference_work_effective_treatment* have the most significant impact on the output of the model. For example, persons who sought treatment or therapy are highly associated with positive predictions, as evidenced by high SHAP scores. Similarly, greater values of *Past_mh_disorder* and *Current_mh_disorder* contribute favourably to the anticipated mental health outcome, emphasising the significance of prior and continuing mental health disorders. Workplace-specific features or characteristics, such as *Role_tech* and

Employer_importance_mental, illustrate the significant impact of occupational contexts on predictions, while *Company_size* reinforces the importance of organisational context. Contextual variables such as *Anonymity_protected* and *Reveal_to_coworkers* are also significant, highlighting the importance of sharing mental health issues at work. Demographic variables such as *Age* and *Country_of_residence_encoded* have a small but detectable influence on the model's predictions.

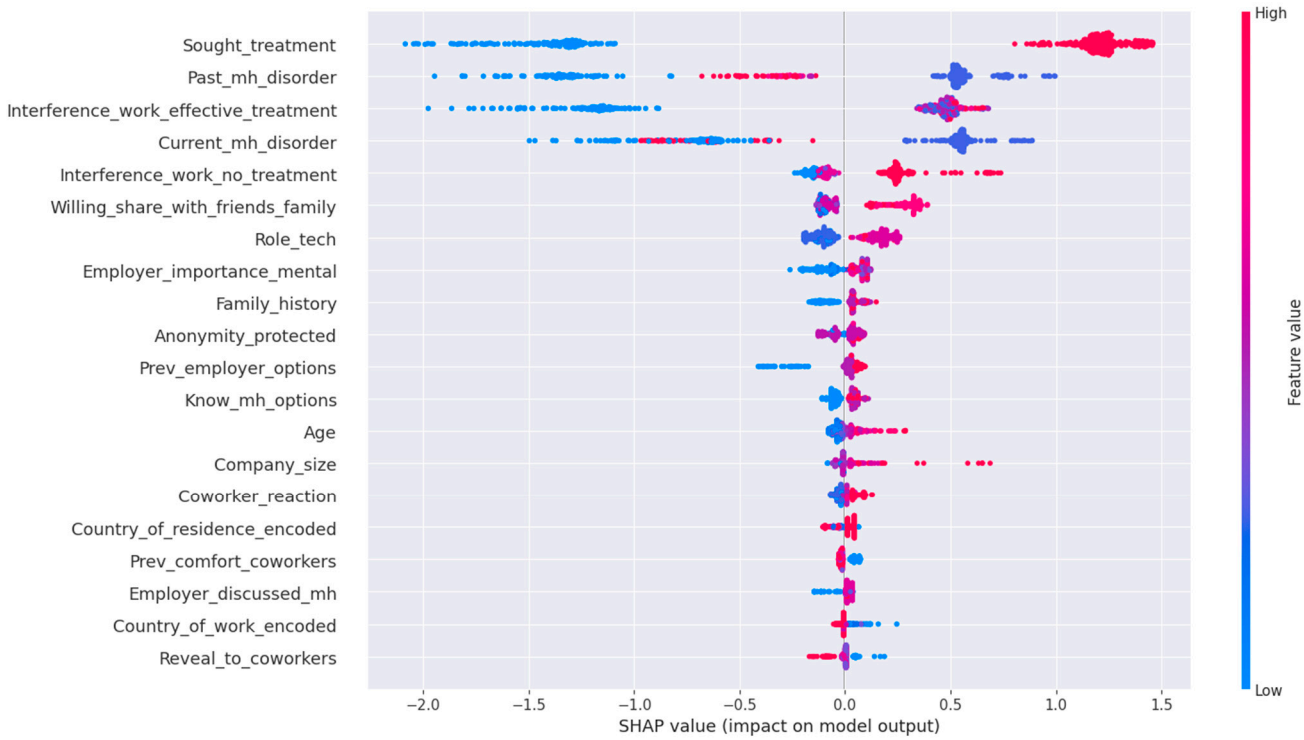


Figure 11. SHAP's Beeswarm summary plot for xGBoost predictions.

The SHAP (Beeswarm) summary plot provides a global explanation of feature contributions, showing both the magnitude and direction (positive or negative) of each predictor's influence on model outcomes across the dataset (Figure 11) [53]. The SHAP values break down the prediction into contributions from individual characteristics, enabling us to see how specific predictors influence the model's output. Features such as *Sought_treatment* and *Current_mh_disorder* had the most significant positive impact, demonstrating the model's emphasis on actionable qualities. With its colour gradient from blue (low feature values) to red (high feature values), the SHAP summary figure visually communicates that greater values of these essential predictors strongly correlate with positive predictions. Overall, the SHAP analysis emphasises the importance of treatment-seeking behaviour, mental health history, and workplace-specific factors in predicting mental health outcomes while providing helpful insight into the xGBoost model's decision-making process. This interpretability is critical for understanding the drivers of model predictions and generating actionable insights for mental health interventions.

The force plot in Figure 12 shows how various attributes contribute to predicting the outcome for the dataset's first instance, complementing the SHAP summary plot (Figure 11). Positive contributors, such as *Current_mh_disorder* = *4* and *Past_mh_disorder* = *4*, drive the model's prediction closer to the observed result. In contrast, *Sought_treatment* = *0* and *Interference_work_effective_treatment* = *2* contribute negatively, decreasing the likelihood of the projected mental health outcome. This visualisation is consistent with the findings in Figure 11, which identified variables such as *Sought_treatment*, *Current_mh_disorder*, and *Past_mh_disorder* as crucial to the model's decision-making processes.

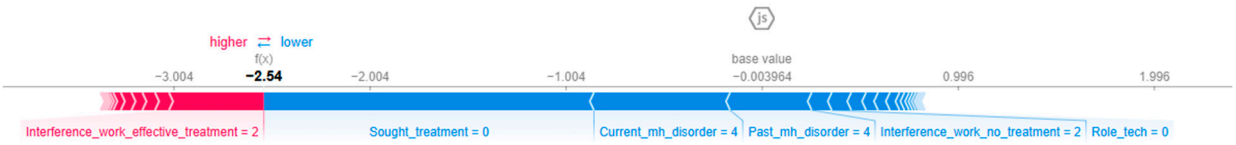


Figure 12. SHAP force plot.

The force plot further breaks down each feature's localised impact for this instance, demonstrating how both positive and negative factors contribute to the prediction. Notably, workplace-related features, such as *Interference_work_effective_treatment*, play an important role, supporting the earlier observation that occupational features are critical in predicting mental health outcomes. The similarity between Figures 11 and 12 demonstrates the model's dependence on important predictors, including treatment-seeking behaviour, mental health status, and workplace-related variables, when generating predictions. Such precise visualisations provide transparent and interpretable insights into the model's logic, facilitating the identification of actionable elements for improving mental health outcomes.

Figure 13 shows a bar plot with the average SHAP values for all features, indicating their overall impact on the xGBoost model predictions. Features with the highest mean SHAP values, such as *Sought_treatment*, *Past_mh_disorder*, and *Interference_work_effective_treatment*, dominate the predictions, demonstrating their importance. *Current_mh_disorder* also makes a significant contribution, emphasising the importance of mental health history and treatment-seeking behaviour in predicting mental health outcomes. *Role_tech* and *Interference_work_no_treatment* offer moderate contributions, but predictors such as *Anonymity_protected* and *Company_size* have more minor but significant impacts.

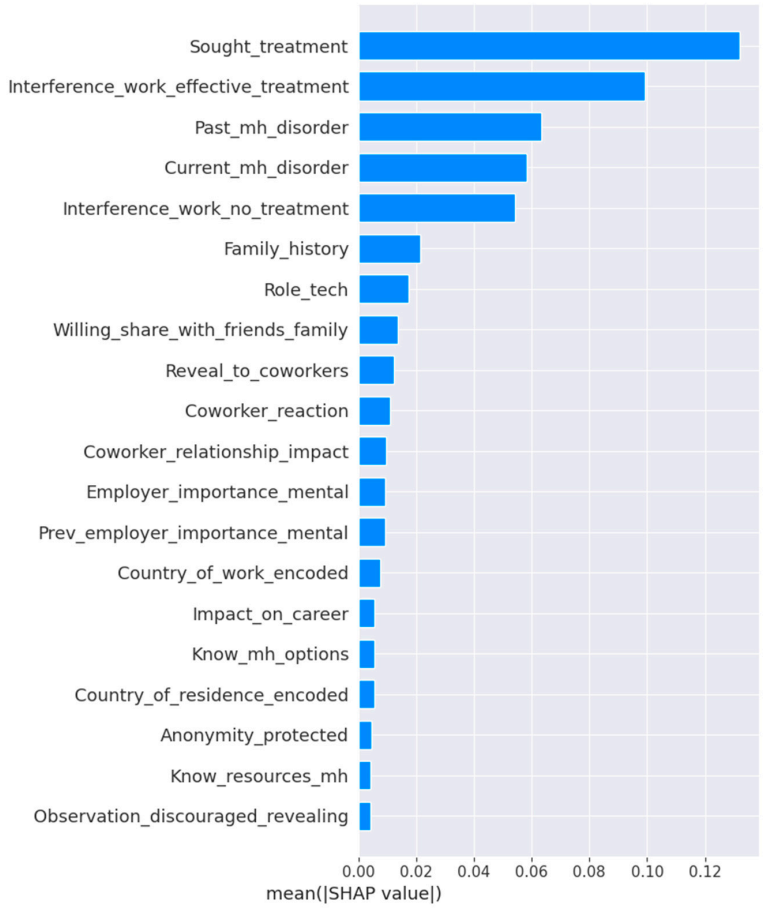


Figure 13. SHAP's bar plot for random forest predictions.

These results are consistent with the SHAP summary plot in Figure 11, which identified *Sought_treatment*, *Past_mh_disorder*, and *Current_mh_disorder* as the most influential predictors. Furthermore, the force map for the first occurrence (Figure 12) confirms these findings, indicating *Sought_treatment* and *Current_mh_disorder* as significant contributors to the model's prediction. The consistency of these visualisations shows that treatment-seeking conduct, past mental health disorders, and workplace-specific characteristics all have a significant impact on the model's decision-making process. This boosts confidence in the model's predictions and provides actionable information for enhancing mental health support and workplace situations.

4.2.2. LIME Analysis

In this study, LIME was applied to the AdaBoost model, since SHAP does not provide direct support for AdaBoost. Attempts to use SHAP's *KernelExplainer* failed due to compatibility issues. As a result, LIME was used as an alternative to ensure AdaBoost's interpretability. A LIME explanation was generated using LIME, incorporating the training data, feature names, target class, and binary classification mode. For explanation, one occurrence from the dataset was chosen, specifically the first instance, to ensure consistency with the SHAP process. The LIME explanation visualised the most relevant workplace-specific features influencing the AdaBoost model predictions for the selected instance. This visualisation provided a clear understanding of the AdaBoost model's decision-making process, supporting the findings from SHAP for xGBoost.

The LIME explanation in Figure 14 provides a clear breakdown of the AdaBoost model's predictions for the dataset's first instance, with a 68% chance of Class 0 (not diagnosed with a mental health disorder) and a 32% chance of Class 1. The attribute *Sought_treatment (value = 0)* has the most significant negative influence, significantly lowering the probability of Class 1. This finding is consistent with its dominant role in the SHAP summary plot (Figure 11) and force plot (Figure 12), where treatment-seeking conduct was a key predictor of the model's predictions. In contrast, features such as *Interference_work_effective_treatment (value = 2)*, *Past_mh_disorder (value = 4)*, and *Current_mh_disorder (value = 4)* make modest positive contributions, raising the probability of Class 1. This is consistent with SHAP findings, which highlight the impact of prior and current mental health disorders on predictions. In addition, features relating to job productivity (*Productivity_affected = 1*) and comfort in disclosing mental health difficulties (*Reveal_to_clients = 1*) made smaller contributions, indicating that they were less influential than major predictors.

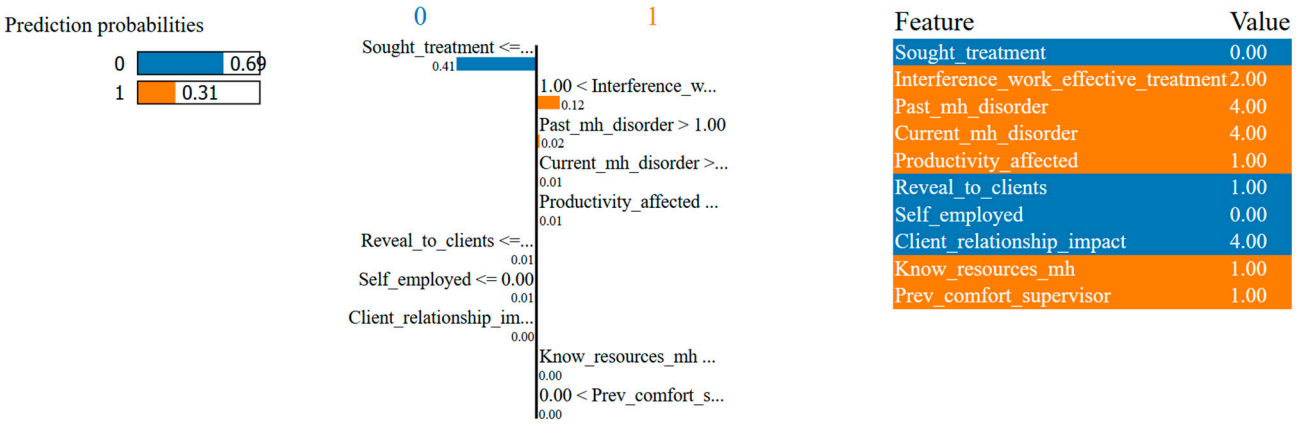


Figure 14. LIME explanation plot for AdaBoost.

The LIME explanation emphasises the localised impact of individual features for this specific instance, which is consistent with the global trends observed in the SHAP analysis. The alignment between the two techniques demonstrates the importance of treatment-seeking behaviour and mental health history as the main predictors of the AdaBoost model. LIME supplements the global insights of SHAP by providing instance-level interpretability and increasing transparency and trust in the decision-making process of the model, particularly for workplace mental health applications. This level of information is critical for stakeholders who want to find actionable features and successfully prioritise efforts.

The local explanation bar graph in Figure 15 shows how the AdaBoost model predicted Class 1 (diagnosed with a mental health disorder) for the dataset's first instance. The attribute *Sought_treatment (value* $\leq$ *0.00)* has the highest negative influence, significantly decreasing the likelihood of Class 1. This observation is consistent with the findings of the SHAP summary plot (Figure 11) and the force plot (Figure 12). The feature *Interference_work_effective_treatment (value* $>$ *1.00 and* $\leq$ *2.00)* has the highest positive contribution, boosting the chance of Class 1. Other features, such as *Past_mh_disorder* $>$ *1.00* and *Current_mh_disorder* $>$ *1.00*, contribute positively, even though to a lower extent, suggesting their moderate roles, as reported in earlier SHAP analyses. *Productivity_affected* and *Reveal_to_clients* have minor positive impacts, but *Self_employed* and *Know_resources_mh* contribute minimally to the prediction. The localised explanation provided by LIME is consistent with the global SHAP results, highlighting the importance of treatment-seeking behaviour and mental health history as the main drivers of the model's predictions.

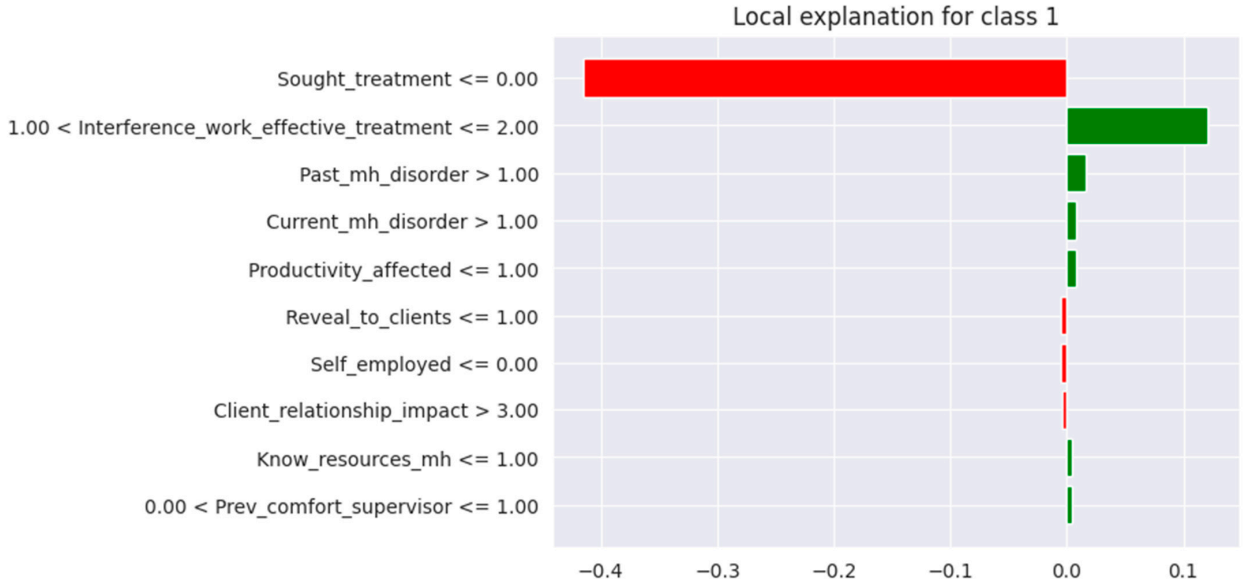


Figure 15. LIME's local explanation bar graph for a single (first) instance.

By providing a clear instance-level breakdown of feature implications, this explanation helps mental health experts to comprehend the decision-making logic of the model. The bar graph depicts how essential features interact to produce the prediction, while the consistency of the model with global SHAP findings increases confidence in its robustness and interpretability. Such findings are critical for increasing trust in the model's applicability in workplace mental health settings and providing opportunities for designing interventions.

These combined efforts of SHAP and LIME reaffirm the importance of explainable and interpretable AI in workplace mental health predictions. These strategies enhance the models' transparency, trustworthiness, and practical application by identifying key workplace-specific aspects, including seeking treatment and current mental health conditions, and providing explicit insights into their impact on predictions. Unlike traditional

black-box approaches, SHAP and LIME enable stakeholders, such as HR experts and decision-makers, to understand why the model made specific predictions. This information enables organisations to make informed decisions about improving mental health support systems and implementing targeted interventions. These explainable models bridge the gap between predictive performance and interpretability, ensuring that AI models can be trusted and effectively implemented in sensitive workplace settings.

5. Discussion

The study's results demonstrate that ensemble-based methods, particularly xGBoost and RF, achieved the highest accuracy (91%) and consistent cross-validation performance (94%), outperforming linear and margin-based models such as SVM and AdaBoost. These findings highlight the suitability of boosted tree models for capturing the non-linear interactions and categorical–ordinal feature mix typical of the OSMI dataset.

The results are consistent with previous research on the OSMI dataset and related workplace mental health prediction tasks. For example, Lohia et al. [13] reported that ensemble-based approaches outperformed baseline classifiers, achieving an accuracy of 90.5%, which is comparable to the results achieved in this study. Similarly, Almaleh [17] applied multiple classifiers on the OSMI survey and found that AdaBoost and xGBoost achieved the highest accuracy (85%), closely aligning with the superior performance of boosted ensembles in this work. Sujal et al. [25] also observed competitive results for KNN and xGBoost (84–85%), further confirming the strength of tree-based models. Kapoor and Goel [26] compared the 2014 and 2021 OSMI datasets and reported peak accuracies, with DT (82%) and AdaBoost (81%) performing strongly, echoing the results of this study, where ensemble models led to better performance. In addition, Mallick and Panda [24] demonstrated that feature selection enhances the predictive power of RF, yielding 78% accuracy compared to other classifiers in the study, which supports this study's finding that RF remains a strong baseline for workplace-focused tasks. Earlier studies, such as Srividya et al. [23], also demonstrated the effectiveness of ensemble methods (90% accuracy), highlighting the broader trend of tree-based algorithms excelling in mental health classification tasks. Finally, Rautaray et al. [22] achieved 73% using DT and RF, reinforcing the consistency of this study's performance outcomes with the literature.

Building on these performance results, this study advances the literature by integrating XAI into workplace-focused models. While prior studies largely prioritised predictive accuracy, they offered limited interpretability. The results of this study further show the effectiveness of XAI techniques, specifically SHAP and LIME, in achieving both high model performance and interpretability for workplace mental health predictions. The SHAP analysis identified crucial predictors, such as *Sought_treatment* and *Current_mh_disorder*, which consistently influenced model decisions across several visualisations, highlighting their importance in assessing the mental health state of employees. These findings highlight the significance of treatment-seeking behaviour and present mental health issues as actionable indicators that can help guide focused therapies. Similarly, LIME provides additional, instance-specific insights, emphasising the importance of workplace-specific aspects such as *interference_work_effective_treatment* and mental health productivity impacts. These XAI approaches provide a strong, interpretable foundation for comprehending global and local predictions.

This study expands on previous research on workplace mental health using ML and the OSMI dataset by filling a major gap: the lack of interpretability in ML models. Prior research that employed the OSMI dataset, such as that aimed at identifying mental health risk factors or predicting mental health outcomes, mostly focused on improving accuracy and other performance measures while failing to address the need for explainability [13,17,24,25].

While these studies have demonstrated the predictive power of ML in workplace mental health situations, their reliance on black-box models has limited the practical utility of their findings for stakeholders seeking actionable insights. In contrast, this study used XAI techniques (SHAP and LIME) to improve model decision-making transparency. By identifying significant predictors such as *Sought_treatment* and *Current_mh_disorder*, the findings contribute to past research by providing a deeper understanding of how individual factors influence mental health predictions. This is a huge step forward, as organisational leaders can make informed decisions based on interpretable models, bridging the gap between predictive performance and real-world applicability.

Furthermore, the findings challenge previous studies that overlooked interpretability by emphasising the relevance of practical insights for building workplace interventions. The use of XAI approaches validates the importance of predictors established in prior studies while also emphasising their distinctive contributions to individual predictions, resulting in an understanding of mental health outcomes. As a result, this study adds to the existing body of knowledge by emphasising the need for accuracy and explainability, ensuring that ML models can anticipate and inform effective mental health measures in the workplace.

5.1. Contributions and Implications of the Study

This study bridges a critical gap in workplace mental health research by combining XAI methodologies with ML models. Previous studies that used ML for mental health prediction lacked transparency and interpretability. This makes it difficult for stakeholders to derive practical insights and integrate mental health ML models into real practice. This study sought to bridge the gap between predictive accuracy and model explainability, ensuring that workplace-specific factors influencing mental health outcomes were discovered and understood. Throughout the process, this study makes theoretical contributions and has theoretical and practical implications.

5.1.1. Theoretical Contributions

This study contributes to the theoretical understanding of mental health prediction by utilising XAI techniques such as SHAP and LIME to identify the most influential workplace-specific indicators. It builds on previous literature by demonstrating how interpretable models can uncover complex relationships between variables and mental health outcomes. Using multi-year data, this study advances theoretical perspectives on predicting mental health outcomes within evolving workplace contexts.

5.1.2. Theoretical Implications

The application of XAI approaches to mental health prediction emphasises the importance of moving beyond black-box models. The study demonstrates that theoretical frameworks in workplace mental health research should incorporate transparency and interpretability to ensure that results are both scientifically sound and practical. This study shows the importance of understanding what predicts mental health outcomes and *how* and *why* these predictors are relevant, thereby providing a basis for future studies to build on interpretable frameworks.

5.1.3. Practical Implications

The practical implications of this study lie in its actionable insights for improving workplace mental health management. By identifying influential predictors such as treatment-seeking behaviour and current mental health status, organisations can design targeted wellness programmes that directly address job stressors and barriers to care. These insights support the development of stigma-free policies, enhanced access to mental health

resources, and personalised support mechanisms, including tailored interventions for employees constrained by job demands. In addition, SHAP and LIME outputs provide decision-makers with clear guidance for prioritising resources towards high-risk employee groups, improving the efficiency and effectiveness of intervention strategies. The use of interpretable models has the potential to enhance transparency and stakeholder trust, bridging the gap between ML predictions and their application in real-world workplace settings.

These insights enables the development of targeted interventions, the design of personalised support mechanisms, and the effective addressing of policies. This study offers a coherent combination of predictive accuracy and interpretability, thereby facilitating the better application of AI in mental health conditions and demonstrating its practical and ethical integration in workplaces. In addition, this study contributes to achieving multiple Sustainable Development Goals (SDGs), such as SDG 3: Good Health and Well-being, which promotes mental health and well-being; SDG 8: Decent Work and Economic Growth, which promotes supportive and productive work environments, and SDG 9: Industry, Innovation, and Infrastructure, which promotes the innovative use of AI in workplace management. By aligning with these objectives, the study emphasises the significance of human-centred, ethical AI solutions that drive both organisational and societal advancement.

5.2. Limitations of the Study

This study has several limitations. The dataset includes tech-focused workplaces, which may limit the applicability of findings to other industries or demographics with different workplace dynamics. The secondary dataset also merged pre- and post-COVID responses, which may reflect different patterns of mental health outcomes. The study intended to capture long-term and diverse conditions; future research could benefit from stratified analyses that separately evaluate pre- and post-COVID data. SMOTE was used as the sole resampling technique to address the class imbalance without exploring alternative approaches, which may have influenced the results. Similarly, data imputation relied solely on the MICE method without considering alternative imputation techniques that could have produced different results. While successful, these options may limit the identification of more efficient preprocessing approaches.

5.3. Future Work

Future research should aim to expand the dataset to include diverse industries and demographics, thereby improving the generalisability of models. Exploring alternative resampling (such as borderline SMOTE or hybrid under-oversampling) and imputation strategies may enhance preprocessing efficiency and model performance. Future work should also empirically compare black-box and explainable approaches to further explore the balance between accuracy and interpretability. In addition, employing longitudinal data for analysis could help track changes in mental health risk over time, enabling proactive and adaptive interventions. Incorporating deep learning models, capable of handling complex data structures, may improve prediction accuracy. Given the inherent complexity of such models, the integration of advanced XAI techniques will be crucial, as these methods can provide transparent, interpretable insights into model decisions. Although SHAP and LIME provided interpretable outputs in this study, they may not fully capture the different requirements of healthcare users.

6. Conclusions

This study highlights the transformative potential of integrating XAI techniques into ML models for predicting workplace mental health. This study addressed key challenges in applying AI to sensitive domains such as mental health by achieving a balanced trade-off

between predictive accuracy and interpretability. SHAP and LIME enabled global and local interpretability, providing stakeholders with insights into the features that drive model decisions and making these insights accessible and actionable.

One of the most notable contributions of the study is its ability to identify workplace-specific characteristics that influence mental health outcomes, allowing for the development of targeted interventions. The findings highlight the need to use explainable models to foster trust and transparency, ensuring that predictions are accurate and understandable to organisational leaders. These models enable organisations to make more informed decisions, to implement effective mental health programmes, and build supportive workplace cultures.

This study also used multi-year data, which provided a unique perspective on changing mental health trends and the effect of workplace dynamics over time. This technique enhances the generalisability of the findings and underscores the importance of tailoring interventions to changing situations. Furthermore, incorporating XAI into complex ML frameworks, such as boosting classifiers and vector machines, demonstrates the viability of using ML models to predict workplace mental health.

This study contributes to the growing body of literature advocating for transparency by enhancing the academic understanding of interpretable AI and its potential practical applications in the workplace to support mental health initiatives. Finally, this study demonstrates the power of XAI to provide explanations of the model's predicted outcomes.

Author Contributions: Conceptualization, T.M.; methodology, T.M., T.B. and E.M.; software, T.M.; validation, T.B. and E.M.; formal analysis, T.M., T.B. and E.M.; investigation, T.M.; resources, T.M.; data curation, T.M.; writing—original draft preparation, T.M.; writing—review and editing, T.B. and E.M.; visualisation, T.M.; supervision, T.B. and E.M.; project administration, T.B.; funding acquisition, T.B. and E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Department of Applied Information Systems Research Ethics Committee (AIS REC) at the University of Johannesburg (ethical clearance code: 2023AIS009).

Informed Consent Statement: Patient consent was waived because the study used de-identified, publicly available secondary data from the Open Sourcing Mental Illness (OSMI) project.

Data Availability Statement: The data used in this study is openly available from the Open Sourcing Mental Illness (OSMI) project at https://osmihelp.org/research.html (accessed on 4 July 2025). No new data were created or generated during this study.

Acknowledgments: The authors would like to express their special thanks to the Department of Applied Information Systems at the University of Johannesburg for their support throughout this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
SMOTE	Synthetic Minority Oversampling Technique
XAI	Explainable Artificial Intelligence

References

1. Dybdahl, R.; Lien, L. Mental health is an integral part of the sustainable development goals. *Prev. Med. Community Health* **2018**, *1*, 1–3. [CrossRef]
2. LaMontagne, A.D.; Martin, A.; Page, K.M.; Reavley, N.J.; Noblet, A.J.; Milner, A.J.; Keegel, T.; Smith, P.M. Workplace mental health: Developing an integrated intervention approach. *BMC Psychiatry* **2014**, *14*, 131. [CrossRef]
3. World Health Organization. *WHO Guidelines on Mental Health at Work*; World Health Organization: Geneva, Switzerland, 2022.
4. Bryan, M.L.; Bryce, A.M.; Roberts, J. The effect of mental and physical health problems on sickness absence. *Eur. J. Health Econ.* **2021**, *22*, 1519–1533. [CrossRef] [PubMed]
5. Mokheleli, T. A Comparison of Machine Learning Techniques for Predicting Mental Health Disorders. 2023. Available online: https://hdl.handle.net/10210/511142 (accessed on 29 October 2024).
6. Cuijpers, P. The Dodo Bird and the need for scalable interventions in global mental health—A commentary on the 25th anniversary of Wampold et al. (1997). *Psychother. Res.* **2023**, *33*, 524–526. [CrossRef] [PubMed]
7. Peykar, P. The Relationship Between Servant Leadership, Supportive Work Environment, and Tech Employees' Mental Well-Being. 2024. Available online: https://digitalcommons.liberty.edu/doctoral/6027 (accessed on 9 December 2024).
8. Bhavya, C. Impact of Mental Health Initiatives on Employee Productivity. *Int. J. Sci. Res. Eng. Manag.* **2025**. Available online: https://doi.org/10.55041/IJSREM46519 (accessed on 5 July 2025).
9. Celestin, P.; Vanitha, N. Mental Health Matters: How Companies Are Failing Their Employees. *SSRN Electron. J.* **2024**, *8*, 44–52. [CrossRef]
10. Özer, G.; Escartín, J. Imbalance between Employees and the Organisational Context: A Catalyst for Workplace Bullying Behaviours in Both Targets and Perpetrators. *Behav. Sci.* **2024**, *14*, 751. [CrossRef] [PubMed]
11. Memish, K.; Martin, A.; Bartlett, L.; Dawkins, S.; Sanderson, K. Workplace mental health: An international review of guidelines. *Prev. Med.* **2017**, *101*, 213–222. [CrossRef]
12. Balch-Samora, J.; Van Heest, A.; Weber, K.; Ross, W.; Huff, T.; Carter, C. Harassment, Discrimination, and Bullying in Orthopaedics: A Work Environment and Culture Survey. *J. Am. Acad. Orthop. Surg.* **2020**, *28*, e1097–e1104. [CrossRef]
13. Lohia, A.; Ranjith, A.; Sachdeva, A.; Bhat, A. An Ensemble Technique to Analyse Mental Health of Workforce in the Corporate Sector. In Proceedings of the 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 17–19 May 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1245–1251. [CrossRef]
14. Graham, S.; Depp, C.; Lee, E.E.; Nebeker, C.; Tu, X.; Kim, H.-C.; Jeste, D.V. Artificial Intelligence for Mental Health and Mental Illnesses: An Overview. *Curr. Psychiatry Rep.* **2019**, *21*, 116. [CrossRef]
15. Iyortsuun, N.K.; Kim, S.-H.; Jhon, M.; Yang, H.-J.; Pant, S. A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. *Healthcare* **2023**, *11*, 285. [CrossRef]
16. Mokheleli, T.; Bokaba, T.; Museba, T. An In-Depth Comparative Analysis of Machine Learning Techniques for Addressing Class Imbalance in Mental Health Prediction. 2023. Available online: https://aisel.aisnet.org/acis2023/15 (accessed on 12 July 2025).
17. Almaleh, A. Machine Learning-Based Forecasting of Mental Health Issues Among Employees in the Workplace. In Proceedings of the 2023 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), Bali, Indonesia, 13–15 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 118–124. [CrossRef]
18. Chauhan, T.; Renjith, P. Utilizing Machine Learning to foster employee mental health in modern workplace environment. In Proceedings of the 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 24–25 February 2024; IEEE: Piscataway, NJ, USA; 2024; pp. 1–5. [CrossRef]
19. Frasca, M.; La Torre, D.; Pravettoni, G.; Cutica, I. Explainable and interpretable artificial intelligence in medicine: A systematic bibliometric review. *Discov. Artif. Intell.* **2024**, *4*, 15. [CrossRef]
20. Nanath, K.; Balasubramanian, S.; Shukla, V.; Islam, N.; Kaitheri, S. Developing a mental health index using a machine learning approach: Assessing the impact of mobility and lockdown during the COVID-19 pandemic. *Technol. Forecast. Soc. Change* **2022**, *178*, 121560. [CrossRef]
21. Joyce, D.W.; Kormilitzin, A.; Smith, K.A.; Cipriani, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digit. Med.* **2023**, *6*, 6. [CrossRef]
22. Rautaray, S.S.; Nayak, S.; Pandey, M. A Machine Learning Based Employee Mental Health Analysis Model. In Proceedings of the 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 15–17 November 2023; IEEE: Piscataway, NJ, USA; 2023; pp. 1055–1059. [CrossRef]
23. Srividya, M.; Mohanavalli, S.; Bhalaji, N. Behavioral Modeling for Mental Health using Machine Learning Algorithms. *J. Med. Syst.* **2018**, *42*, 88. [CrossRef]
24. Mallick, S.; Panda, M. Predictive Modeling of Mental Illness in Technical Workplace: A Feature Selection and Classification Approach. In Proceedings of the 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, Raigarh, India, 5–7 June 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–7. [CrossRef]

25. Sujal, B.; Neelima, K.; Deepanjali, C.; Bhuvanashree, P.; Duraipandian, K.; Rajan, S.; Sathiyanarayanan, M. Mental Health Analysis of Employees using Machine Learning Techniques. In Proceedings of the 2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS), Bangalore, India, 4–8 January 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
26. Kapoor, A.; Goel, S. Predicting Stress at Workplace using Machine Learning Techniques. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5. [CrossRef]
27. Chung, J.; Teo, J. Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges. *Appl. Comput. Intell. Soft Comput.* **2022**, *2022*, 1–19. [CrossRef]
28. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2020**, *23*, 18. [CrossRef]
29. Zilker, S.; Weinzierl, S.; Zschech, P.; Kraus, M.; Matzner, M. Best of Both Worlds: Combining Predictive Power with Interpretable and Explainable Results for Patient Pathway Prediction. In Proceedings of the ECIS 2023, Kristiansand, Norway, 11–16 June 2023; Available online: https://aisel.aisnet.org/ecis2023_rp/266 (accessed on 9 December 2024).
30. Dwivedi, K.; Rajpal, A.; Rajpal, S.; Kumar, V.; Agarwal, M.; Kumar, N. Enlightening the path to NSCLC biomarkers: Utilizing the power of XAI-guided deep learning. *Comput. Methods Programs Biomed.* **2024**, *243*, 107864. [CrossRef]
31. Letoffe, O.; Huang, X.; Asher, N.; Marques-Silva, J. From SHAP Scores to Feature Importance Scores. *arXiv* **2024**, arXiv:2405.11766. [CrossRef]
32. Mienye, I.D.; Obaido, G.; Jere, N.; Mienye, E.; Aruleba, K.; Emmanuel, I.D.; Ogbuokiri, B. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. *Inform. Med. Unlocked.* **2024**, *51*, 101587. [CrossRef]
33. Noor, A.A.; Manzoor, A.; Mazhar Qureshi, M.D.; Qureshi, M.A.; Rashwan, W. Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions. *WIREs Data Min. Knowl. Discov.* **2025**, *15*, e70018. [CrossRef]
34. Atlam, E.S.; Rokaya, M.; Masud, M.; Meshref, H.; Alotaibi, R.; Almars, A.M.; Assiri, M.; Gad, I. Explainable artificial intelligence systems for predicting mental health problems in autistics. *Alex. Eng. J.* **2025**, *117*, 376–390. [CrossRef]
35. Ahmed, I.; Brahmacharimayum, A.; Ali, R.H.; Khan, T.A.; Ahmad, M.O. Explainable AI for Depression Detection and Severity Classification From Activity Data: Development and Evaluation Study of an Interpretable Framework. *JMIR Ment. Health* **2025**, *12*, e72038. [CrossRef] [PubMed]
36. OSMI: About, O.S.M.I. Available online: https://osmihelp.org/about/about-osmi.html (accessed on 1 December 2024).
37. Kelloway, E.K.; Dimoff, J.K.; Gilbert, S. Mental Health in the Workplace. *Annu. Rev. Organ. Psychol. Organ. Behav.* **2023**, *10*, 363–387. [CrossRef]
38. Alruhaymi, A.Z.; Kim, C.J. Why Can Multiple Imputations and How (MICE) Algorithm Work? *Open J. Stat.* **2021**, *11*, 759–777. [CrossRef]
39. International Labour Organization. *ILO Convention No. 138 at a Glance*; International Labour Organization: Geneva, Switzerland, 2018; Available online: https://www.ilo.org/publications/ilo-convention-no-138-glance (accessed on 10 December 2024).
40. Bichri, H.; Chergui, A.; Hain, M. Investigating the Impact of Train/Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 331–339. [CrossRef]
41. Komorowski, M.; Marshall, D.C.; Salciccioli, J.D.; Crutain, Y. Exploratory Data Analysis. In *Secondary Analysis of Electronic Health Records*; Springer International Publishing: Cham, Switzerland, 2016; pp. 185–203.
42. Alghamdi, M.; Al-Mallah, M.; Keteyian, S.; Brawner, C.; Ehrman, J.; Sakr, S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. *PLoS ONE* **2017**, *12*, e0179805. [CrossRef]
43. Sowjanya, A.M.; Mrudula, O. Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. *Appl. Nanosci.* **2023**, *13*, 1829–1840. [CrossRef]
44. Mohamed, E.S.; Naqishbandi, T.A.; Bukhari, S.A.C.; Rauf, I.; Sawrikar, V.; Hussain, A. A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms. *Healthc. Anal.* **2023**, *3*, 100185. [CrossRef]
45. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv* **2016**, arXiv:1603.02754v3.
46. Liu, H.; Yan, G.; Duan, Z.; Chen, C. Intelligent modeling strategies for forecasting air quality time series: A review. *Appl. Soft Comput.* **2021**, *102*, 106957. [CrossRef]
47. Thölke, P.; Mantilla-Ramos, Y.-J.; Abdelhedi, H.; Maschke, C.; Dehgan, A.; Harel, Y.; Kemtur, A.; Mekki Berrada, L.; Sahraoui, M.; Young, T.; et al. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *Neuroimage* **2023**, *277*, 120253. [CrossRef] [PubMed]
48. Padilla, R.; Netto, S.L.; da Silva, E.A.B. A Survey on Performance Metrics for Object-Detection Algorithms. In Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), Niterói, Brazil, 1–3 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 237–242. [CrossRef]
49. Aurelio, Y.S.; de Almeida, G.M.; de Castro, C.L.; Braga, A.P. Cost-Sensitive Learning based on Performance Metric for Imbalanced Data. *Neural Process Lett.* **2022**, *54*, 3097–3114. [CrossRef]

50. Ma, X.; Hou, M.; Zhan, J.; Liu, Z. Interpretable Predictive Modeling of Tight Gas Well Productivity with SHAP and LIME Techniques. *Energies* **2023**, *16*, 3653. [CrossRef]
51. Wong, T.-T.; Yeh, P.-Y. Reliable Accuracy Estimates from *k*-Fold Cross Validation. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 1586–1594. [CrossRef]
52. Marcilio, W.E.; Eler, D.M. From explanations to feature selection: Assessing SHAP values as feature selection mechanism. In Proceedings of the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Recife/Porto de Galinhas, Brazil, 7–10 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 340–347. [CrossRef]
53. Ponce-Bobadilla, A.V.; Schmitt, V.; Maier, C.S.; Mensing, S.; Stodtmann, S. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin. Transl. Sci.* **2024**, *17*, e70056. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.