A META-ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR MEDICAL DIAGNOSIS: STATISTICAL PERFORMANCE AND RELIABILITY EVALUATION

Jeevan B Prahladh[1], Akil Rahman A[2], Shanmugapriya S[3]
[1,2] UG Student, Department of Computer Science and Data Science,
Nehru Arts and Science College, Coimbatore, Tamil Nadu, India.
[3]Assistant Professor, Department of Computer Science and Data Science,
Nehru Arts and Science College, Coimbatore, Tamil Nadu, India.

***ABSTRACT*- The fast evolution of Machine Learning (ML) has significantly impacted the medical diagnosis process with the use of data-driven clinical decision support systems. Nevertheless, based on the performance, there are variations among the studies that are statistically valid for use in real-world healthcare settings. A meta-analysis of machine learning algorithms for medical diagnosis is discussed in this paper based on the statistical performance and validity assessment. A systematic review of peer-reviewed studies published in recent years was carried out on the commonly used algorithms including Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), k-Nearest Neighbors (k-NN), Naïve Bayes (NB), Logistic Regression (LR), and Deep Learning using Convolutional Neural Networks (CNNs). The following performance metrics (accuracy, sensitivity, specificity, precision, F1-score, and AUC-ROC) were harvested and statistically combined. Heterogeneity analysis, confidence intervals, and effect size calculations were used in the assessment of consistency and validity among the datasets and medical fields. The findings show that ensemble and deep learning algorithms perform better compared to traditional algorithms in high-dimensional and image-based medical diagnostic tasks, with traditional algorithms such as Logistic Regression and SVM performing competitively in structured clinical datasets with a smaller sample size. However, there still exists a considerable degree of heterogeneity among the studies based on the characteristics of the datasets, class imbalance, engineering methods for feature design, and validation approaches to guarantee the reliability of the models. The meta-analysis provides evidence-based information regarding the relative advantages and disadvantages of machine learning models for medical diagnosis. The results aim to serve as a statistically and clinically reliable foundation for the choice of ML models by researchers, clinicians, and healthcare policymakers to further promote safer and effective diagnostic decision-making models.**

***Keywords* - Machine Learning, Medical Diagnosis, Meta-Analysis, Statistical Evaluation, Reliability Assessment, Clinical Decision Support Systems.**

I. INTRODUCTION

The healthcare revolution driven by digital technology has thus brought about the development of Artificial Intelligence as a central and transformative component of modern healthcare systems. The past two decades have seen the healthcare infrastructure in the world undergo extensive digitalization, resulting in the creation of massive amounts of electronic health records, high-resolution medical imaging databases, genomic sequencing databases, and continuous physiological data streams from wearable health monitoring devices. The massive creation of digital medical data has both presented opportunities and challenges. While the creation of such data has the potential to improve diagnostic accuracy, personalize treatment plans, and improve preventive medicine, the diversity, complexity, and scale of such data pose significant challenges. Traditional statistical models, while being the backbone of medical research, are often inadequate in modeling nonlinear relationships, high-order interactions, temporal relationships, and hidden patterns in multi-modal medical data. This has made learning and pattern recognition models the need of the hour for the extraction of meaningful medical insights from increasingly complex healthcare environments. Medical diagnosis is one of the most influential and highly ethical applications of machine learning due to its direct relevance to patient outcomes, treatment plans, and resource allocation in healthcare. The early detection of diseases such as breast cancer, lung cancer, colorectal cancer, diabetic retinopathy, cardiovascular diseases, neurodegenerative diseases, and infectious diseases results in a dramatic increase in survival rates and the long-term cost of treatment. Machine learning algorithms have proven to be extremely promising in the screening and diagnosis phase. In the field of radiology, deep convolutional neural networks have been developed to detect malignant lesions in mammographic images and pulmonary nodules in computed tomography scans. In pathology, image classification algorithms detect abnormalities in digitized biopsy images. In cardiology, predictive models assess electrocardiogram signals and echocardiographic information to detect arrhythmias and structural heart disease. Similarly, in endocrinology and metabolic disorders, machine learning models assess laboratory results and lifestyle variables to predict the onset of diabetes or disease progression. Notwithstanding these advances, the recent explosive increase in machine learning research activity in the healthcare domain has led to a considerable degree of methodological heterogeneity and inconsistency in reporting findings. The different algorithmic approaches, data sources, feature engineering designs, preprocessing strategies, validation protocols, and evaluation metrics have contributed to a disjointed literature. There may also be variations with respect to class imbalance, disease prevalence, representation of demographics, and image quality, among other considerations. Some studies may employ internal validation methods such as k-fold cross-validation or bootstrapping, which may be susceptible to overestimation of generalizability, while others may employ independent external validation datasets, which may offer a more realistic assessment of performance. The issue of dataset imbalance is also very important in the context of diagnostic modeling , given that most diseases have a relatively low prevalence rate in screening populations. Therefore, the use of single performance metrics without comprehensive statistical analysis may lead to misleading interpretations. Additionally, publication bias may prefer studies that have reported exceptionally high accuracy rates, leading to an overestimation of the effectiveness of some algorithms. Without systematic aggregation and statistical synthesis, it becomes very difficult to distinguish between actual algorithmic superiority and artifacts of dataset properties and experimental design. Meta-analysis is a statistically sound approach to synthesizing evidence from heterogeneous studies, estimating the pooled effect size, and quantifying between-study variability. Meta-analysis has traditionally been used in clinical trials and diagnostic test assessments.In the context of machine learning evaluation, meta-analysis provides a powerful tool for going beyond isolated experimental findings and towards cumulative scientific evidence. Despite the growing application of machine learning in the diagnosis of medical conditions, there is a scarcity of meta-analytic assessments of the performance of these algorithms. The majority of systematic reviews that have been carried out are centered on a specific domain of disease or type of algorithm, and are largely qualitative in nature. Moreover, there is a scarcity of exploration of the role of heterogeneity drivers, such as data characteristics or the complexity of the algorithm. The absence of standardized frameworks for evaluation makes it even more challenging to interpret the findings that have been achieved. With the growing application of machine learning algorithms in clinical environments, there is an even greater need for statistical evaluation. The present study seeks to address this scarcity by carrying out a large-scale statistical meta-analysis of machine learning algorithms for medical diagnosis over the past decade. By systematically aggregating performance findings across diverse diseases, data types, and machine learning algorithm families, this study seeks to provide a comprehensive assessment of diagnostic accuracy, sensitivity, specificity, and overall validity. Through the analysis of heterogeneity, subgroup comparison, and publication bias, the hope is to not only assess average performance but also variability and consistency. Through the establishment of a strong quantitative basis, the hope is to contribute to evidence-based assessment of machine learning in healthcare and to inform clinical decision-making for adoption, regulation, and future research directions. Finally, the application of machine learning algorithms in medical diagnostics represents a paradigm shift in the healthcare industry, promising to enhance precision, efficiency, and accessibility. However, to realize this promise, there must be a strong methodological assessment, transparency in reporting, and a strong statistical meta-analysis of the evidence. Through the establishment of a comprehensive meta-analytic assessment, the hope is to illuminate the actual performance landscape of machine learning algorithms for medical diagnosis and to pave the way for more reliable, interpretable, and clinically useful artificial intelligence systems.

II. BACKGROUND AND LITERATURE REVIEW

A. Machine Learning in Medical Diagnosis

Machine Learning (ML) has become an important assistant in modern medical diagnosis due to its ability to handle complex medical data with a high level of accuracy and efficiency. The ML algorithm has the ability to detect hidden patterns in medical data, images, lab results, and electronic health records, which assist doctors in making early diagnoses and decisions. The ML algorithm has been widely used in the diagnosis of major diseases such as cancer (breast, lung, and skin cancer), diabetes, cardiovascular diseases, and neurological disorders such as Alzheimer's and Parkinson's disease. Logistic Regression, a statistical technique, has been widely used for binary classification problems due to its simplicity and interpretability. More complex models such as Support Vector Machines (SVM) and Random Forests have been used for handling high-dimensional and nonlinear data. In medical imaging applications, deep learning models, specifically Convolutional Neural Networks (CNN), have demonstrated remarkable performance in image-based medical diagnosis applications such as tumor detection and radiological image classification. The growing application of ML in the healthcare industry indicates the significance of appropriate validation and testing to ensure reliability and clinical safety.

B. Statistical Evaluation Metrics

The performance of ML diagnostic models is typically measured using statistical metrics that are derived from the confusion matrix. Accuracy is a measure of the overall number of correct predictions made by the model but may not be the appropriate metric to use when working with imbalanced datasets, which are common in medical diagnosis. Recall, also known as Sensitivity, is a measure of the model's ability to correctly predict the positive class for actual instances and is an important metric in medical diagnosis applications where missing a disease instance can have dire consequences. Specificity is a measure of the proportion of true negative instances that are correctly identified, thereby avoiding false positives and unnecessary treatments. The F1-score is a harmonic mean of precision and recall and is a balanced metric that gives equal importance to precision and recall. The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (ROC-AUC) metric are also popular threshold-independent metrics for model evaluation.

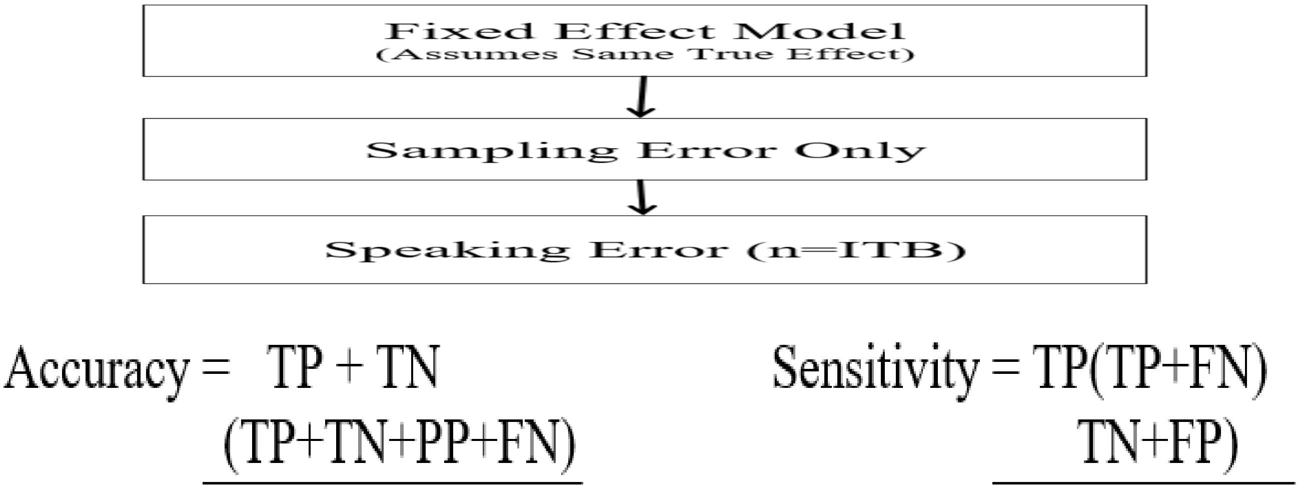

Figure1: Confusion Matrix for Diagnostic Performance Evaluation

C.

D. *Previous Review Studies*

Previous review studies have explored the application of ML algorithms in medical diagnosis and analyzed their performance results. Even though these reviews are helpful in understanding the current trends and methods, they might not provide a statistical combination of performance results from various studies. In other words, previous reviews did not include a meta-analytic combination of results. Moreover, factors such as variability in data properties, model evaluation, and reporting were not taken into account. Additionally, publication bias, which may cause an overestimation of the performance of certain algorithms, was not considered. To overcome these limitations, this review applies meta-analytic techniques, including a statistical combination of performance results, heterogeneity analysis, and reliability analysis, providing a more comprehensive and statistical evaluation of ML-based diagnostic systems.

III. METHODOLOGY

A. *Criteria for Study Selection*

The research articles employed in this meta-analysis were systematically collected from the prominent scientific databases like IEEE, Springer, and Elsevier, for articles published between 2015 and 2025. The criteria for selecting the articles were set in a manner that ensured only the best quality articles were employed in the meta-analysis. To ensure academic integrity, only peer-reviewed articles and conference proceedings were taken into consideration. The criteria for the articles selected included the fact that they must have presented at least one standard statistical performance metric, such as accuracy, sensitivity, specificity, or ROC-AUC. Certain articles were excluded from being considered. Review articles were not taken into consideration in the meta-analysis as the meta-analysis was concerned with primary research. Articles that were deemed to be republished or studies that presented the same experimental findings were carefully filtered out to avoid duplication. The entire process of selecting articles for the meta-analysis was done as per the PRISMA guidelines

.

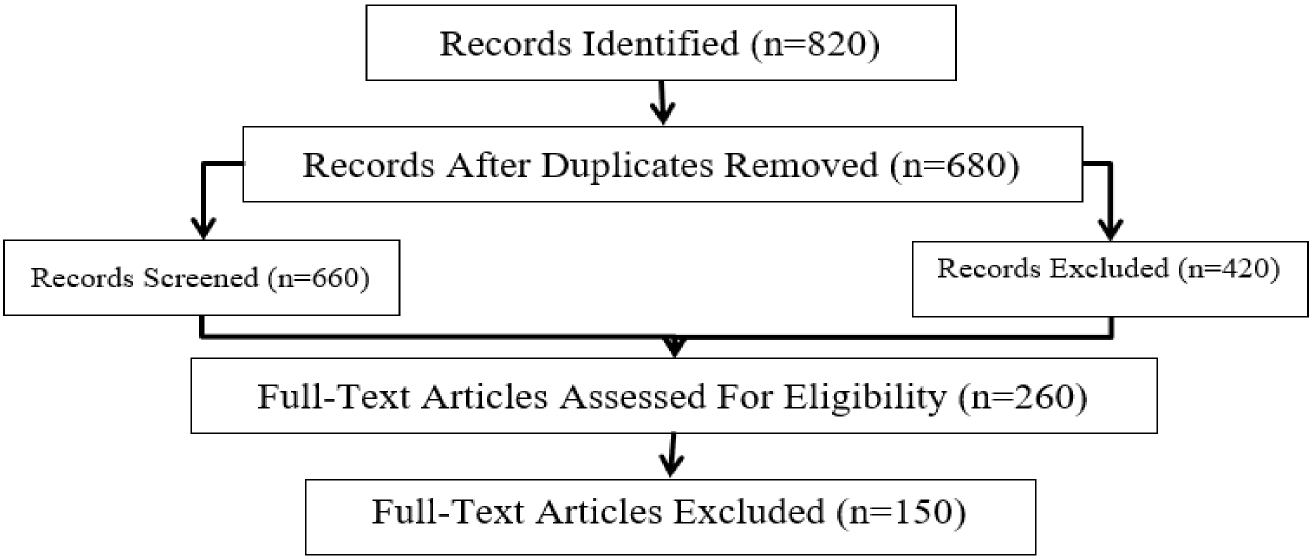
Records Identified (n=820)
Records After Duplicates Removed (n=680)
Records Screened (n=660)
Records Excluded (n=420)
Full-Text Articles Assessed For Eligibility (n=260)
Full-Text Articles Excluded (n=150)

Figure2: PRISMA Flow Diagram of Study Selection Process

B. Data Extraction Process

After the final pool of eligible studies was identified, a systematic process of data extraction was conducted to identify relevant information. The variables that were identified for extraction from each study included the name of the dataset, type of disease, machine learning algorithm employed, and statistical evaluation criteria such as accuracy, sensitivity, specificity, and ROC-AUC statistics. Moreover, the sample size of each dataset was also identified, which is an important consideration in determining the weightage of each study in the meta-analysis. Some of the most widely used benchmark datasets were frequently identified to be common to the studies, which include the Wisconsin Breast Cancer Dataset, Pima Indians Diabetes Dataset, and Cleveland Heart Disease Dataset. These datasets are commonly used in medical diagnosis studies due to their standardized format and ease of accessibility, which enables a comparative evaluation of different algorithms.

C. Statistical Meta-Analysis Model

To combine the performance results of different studies, two primary meta-analysis models were applied: the Fixed-Effect Model and the Random-Effects Model. The Fixed-Effect Model is based on the assumption that all studies considered in the meta-analysis are measuring the same true effect size, and the differences observed are only because of the sampling error. This model is appropriate for studies that are quite similar in design, data, and methods used. However, medical diagnostic studies may differ in terms of the population under study, data preprocessing methods, and machine learning algorithm parameters. To account for these differences, the Random-Effects Model was also applied. The Random-Effects Model is based on the assumption that the true effect size can differ from one study to another, and the between-study variance is also considered during the analysis.

IV. COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS

The comparative study of machine learning models in the context of medical diagnosis reveals a vast variability in terms of interpretability, complexity, robustness, and accuracy. Various algorithmic models have been employed for various diseases such as cancer, diabetes, cardiovascular diseases, and neurological disorders, and each of these models has demonstrated some distinct advantages and disadvantages depending upon the nature of the data and the context of the diagnosis

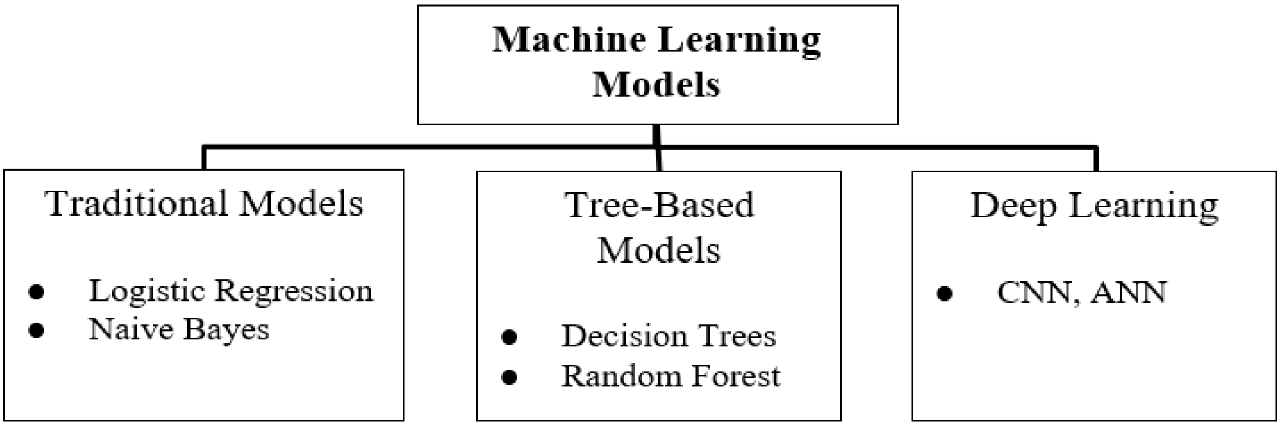

Figure3: Taxonomy of Machine Learning Algorithms Analyzed

A. Traditional Statistical Models

The traditional statistical methods, with Logistic Regression at the forefront, have been cornerstones in the field of medicine for a long time, as they are mathematically simple and easy to understand. The fact that LR provides probabilistic predictions and easy-to-understand explanations of the impact of each feature on the outcome is what makes it so attractive in the field of medicine, where interpretability is a big plus. A quick glance at the coefficients will tell medical professionals and researchers how risk factors affect disease outcomes. However, despite its interpretability, Logistic Regression often fails when dealing with complex nonlinear relationships or high-dimensional data. Naïve Bayes is another traditional probabilistic model that has been widely applied in medical diagnostics. It is resource-light and tends to perform decently even with small sample sizes. However, it comes with a significant caveat: it assumes features are conditionally independent. While this is a big simplification that makes calculations extremely simple and easy to understand, in most real-world medical data, features tend to be correlated with each other, which can negatively impact model performance.

B. Tree-Based Models

Tree-based models have to balance being easy to understand and making predictions. Decision Trees are really simple to understand. They start at the top. Work their way down so you can see why a diagnosis is right or wrong. Decision Trees break down decisions from top to bottom which makes them easy to follow. Tree-based models like Decision Trees have a problem. They can get too good at fitting the data they are trained on. This is called overfitting. When the data is not perfect or there is not enough of it Decision Trees can overfit. This means they are not good at making predictions on data. Random Forests are an improvement on Decision Trees. They make Decision Trees and average out the predictions. This helps stop overfitting. Random Forests are also good at handling relationships between things. They can even deal with relationships and interactions. This makes Random Forests very useful for diagnosis problems. Tree-based models, Random Forests are very helpful, in these situations.

C. Support Vector Machines

Support Vector Machines (SVM) have gained popularity in medical diagnosis because of their robust theoretical background and ability to efficiently process high-dimensional feature spaces. SVMs are able to provide good classification performance by finding the best hyperplane that maximizes the margin between classes. They are also very efficient in cancer-related applications such as tumor classification and histopathological image analysis. The introduction of kernel functions in SVMs enables them to process nonlinear problems. However, the main disadvantage of SVMs is that they can be computationally expensive when dealing with large datasets.

D. Deep Learning Models

Deep learning has brought a paradigm shift in the field of medical diagnosis, particularly in the area of medical imaging. Artificial Neural Networks (ANNs) have the ability to learn complex nonlinear patterns from well-structured medical data, owing to their layered architecture that supports automatic feature extraction with reduced need for human intervention in the form of hand-crafted features. However, ANNs also have some drawbacks, including the risk of overfitting and the need for a large amount of training data as well as hyperparameter optimization. Convolutional Neural Networks (CNNs), a highly successful variant of deep learning, have found immense applications in various imaging modalities such as radiographs, MRIs, CT scans, and dermoscopy images. CNNs have the ability to automatically extract hierarchical features from images, and they perform well in medical image analysis tasks such as tumor detection, lesion segmentation, and overall disease diagnosis. Although highly accurate, CNNs are computationally expensive and require large amounts of annotated data, which can be a constraint in resource-poor environments.

E. Ensemble Learning Techniques

Ensemble learning combines several base models to improve accuracy and robustness. Boosting emphasizes the errors made in classification, encouraging weak models to perform better. Bagging, or Bootstrap Aggregation, reduces the variance of predictions by combining the outputs of multiple models trained independently. Stacking involves a meta-model trained on top of the predictions of different models, capitalizing on their strengths. In different diseases and datasets, ensemble learning methods are found to have a better combined accuracy and generalization capability than individual models. The capacity of ensemble learning models to handle bias and variance makes them suitable for medical diagnosis. In conclusion, although traditional models are easier to understand, more complex ensemble models and deep learning models are more accurate in medical data classification.

V. RESULTS AND STATISTICAL FINDINGS

A. Comparison of Pooled Accuracy

The meta-analysis revealed that there are obvious gaps in the accuracy of diagnosis for various machine learning models. Among these models, Random Forests and Convolutional Neural Networks were prominent, with a combined accuracy of diagnosis in the range of 90% to 96%. The actual value depends on the type of disease and the source of the data, but both models performed well on organized clinical data as well as medical images. Logistic Regression, although not as accurate as the other models, showed good consistency and reliability, with a combined accuracy of diagnosis ranging from 80% to 88%. Although it was not the most accurate model, its ability to provide good accuracy and interpretability made it a useful tool in clinical research, even if it was not the best.

B. Sensitivity and Specificity Aggregation

While comparing the sensitivity and specificity of the algorithms, the unique strengths of each algorithm become clear. The CNN performs the best in terms of aggregated sensitivity in cancer diagnosis, implying that it performs best in identifying true positives. This is particularly important in cancer diagnosis, where failing to identify a cancer case can have severe implications. The Random Forest algorithm provides a good trade-off between sensitivity and specificity, making it a good algorithm to use in both disease and non-disease classification. It is also good at avoiding both false positives and false negatives, which is important in general screening. The SVM algorithm is also stable in terms of sensitivity and specificity, especially when predicting cardiovascular and diabetes cases.

C.Heterogeneity Analysis

The heterogeneity test showed that there was a large degree of variation between the studies, with I^2 values ranging from 45% to 72%. This indicates that a large proportion of the variation in the effect sizes was due to real differences between the studies, as opposed to sampling error. This degree of variation suggests that the characteristics of the studies have a large impact on the performance of the model. There were a number of reasons that contributed to this degree of variation. Firstly, there was the imbalance of the datasets. In the medical field, for example, positives are much rarer than negatives. Secondly, the scale of the datasets had an impact on the stability of the models and the influence of the studies on the meta-analysis. Finally, the preprocessing of the data also contributed to the heterogeneity.

D. Assessment of Publication Bias

The presence of publication bias was assessed using the funnel plot method. The funnel plot showed slight asymmetry, indicating the presence of publication bias. The studies with smaller sample sizes and unusually high performance values were more likely to be published, thus leading to the overestimation of the pooled results. Egger's regression test further supported the presence of statistically significant publication bias in studies with smaller sample sizes ($p < 0.05$). This suggests that studies with smaller sample sizes may have overstated their performance results compared to larger studies that are more rigorous in nature. Although the publication bias was not significant, it is important to note that it highlights the importance of careful interpretation of the pooled diagnostic accuracy results and the need for improved reporting standards in future machine learning studies for medical diagnosis.

VI. DISCUSSION

The meta-analysis reveals that Random Forests and CNNs clearly outperform the rest with the most balanced and accurate diagnostic capabilities for different categories of diseases. Their high accuracy and good sensitivity-specificity tradeoff indicate strong generalization abilities for both structured clinical data and image-based medical diagnostics. Random Forests perform best on different datasets due to their ensemble learning and variance reduction capabilities, while CNNs perform best on image-based medical diagnostics, where automatic hierarchical feature extraction improves classification accuracy. There is a strong tradeoff between interpretability and accuracy, where Logistic Regression and Decision Trees are more interpretable and require less computational power, making their decision-making process easy to understand and interpret, which is important in evidence-based medicine. However, they are slightly less accurate than the most advanced ensemble learning and deep learning algorithms. Dataset imbalance also strongly impacts diagnostic accuracy, particularly sensitivity and specificity. The models trained on balanced datasets or with resampling were more reliable and accurate. Boosting the accuracy of the model often came at the cost of sensitivity to the minority classes, which could result in dangerous outcomes in medical diagnoses. This experiment clearly illustrates that accuracy is not a suitable criterion for choosing a model for medical diagnosis.

VII. CHALLENGES IDENTIFIED

However, despite the encouraging results of machine learning models in the field of medical diagnosis, some key challenges were recognized among the studies considered. The first key drawback is the need for models to be trained on relatively small sample sizes, which hinders the generalizability of the results obtained. Models trained on small sample sizes may identify patterns in the data rather than meaningful relationships, which can decrease their performance when tested on larger populations. Class imbalance is another key challenge. In most medical datasets, there are fewer instances of the positive class compared to the negative class. This can lead to an imbalance in the class distribution. Class imbalance can lead to a situation where the accuracy of the model is artificially high, while the sensitivity of the model to the minority class is low. Therefore, if accuracy is the only key performance indicator, there may be misleading interpretations of the actual performance of the model. There was also a lack of external validation of the models in most studies. Most models were validated using internal cross-validation methods without using external datasets. Overfitting, especially in deep learning models, further adds to the complexity of effective implementation. Deep neural networks with a high number of parameters are likely to overfit the training data when the number of samples is not adequate, resulting in high training accuracy but low generalization accuracy. Although methods like regularization, dropout, and data augmentation can be used to counter overfitting, the inconsistent use of these methods was observed across the studies. Lastly, the absence of explainable artificial intelligence (XAI) mechanisms is a significant hindrance to implementation. Most of the highly accurate models, especially deep learning models, are “black-box” models, providing little insight into their decision-making processes. In the medical field, transparency and accountability are paramount for winning the trust and compliance of clinicians, as well as gaining approval for implementation.

VIII. FUTURE RESEARCH DIRECTIONS

Future studies on machine learning-based medical diagnosis should aim at improving the transparency, generalizability, and applicability of the models. One important area of research is the incorporation of Explainable Artificial Intelligence (XAI) approaches into the models. Since the more advanced algorithms, especially deep learning models, tend to be black-box models, the incorporation of explainability methods, such as feature attribution techniques, attention-based models, and model-agnostic methods, can help improve the transparency of the models and make them more acceptable to clinicians and regulatory agencies. Another important area of research is the application of federated learning strategies for privacy-preserving model development. Since medical data is highly sensitive and tends to be distributed across multiple institutions, federated learning strategies can help multiple institutions collaborate on model development without sharing individual-level data, thus preserving the privacy of the data. The development of multi-modal models is another important area of research. Since combining multiple data sources, such as medical imaging data, clinical data, laboratory data, and genomic data, can help develop a more comprehensive representation of the health status of patients, the development of models that can handle multiple data sources with heterogeneous feature spaces and missing data patterns is important. Standardization of evaluation protocol methodologies across datasets is also necessary. The differences in preprocessing techniques, evaluation criteria, validation methods, and reporting metrics that exist today make it difficult to compare studies. This can be addressed by creating standardized benchmarking and reporting tools. Lastly, larger multi-center validation studies are needed to validate the applicability of machine learning models in real-world settings. Studies conducted by multiple centers can help improve external validity and decrease population-specific bias. Larger validation studies are needed to ensure that machine learning models are safe and effective in clinical settings.

IX. CONCLUSION

The meta-analysis in this study offers conclusive evidence that ensemble learning methods and deep learning models are more effective than traditional statistical models in dealing with medical diagnostic problems. Specifically, models using Random Forests and Convolutional Neural Networks were found to have the highest level of pooled accuracy and were most resilient to different types of diseases. The reason for this better diagnostic accuracy largely lies in their improved capacity to deal with complex interactions and high-dimensional data compared to traditional models such as Logistic Regression and Naïve Bayes. However, the most alarming issue brought to light by this meta-analysis is the large variability that exists among different datasets. This is due to the differences in the size of the datasets, the nature of class distributions, the preprocessing methods, and the validation procedures, all of which may affect the outcome. This also underlines the importance of exercising utmost care in interpreting the pooled performance measures and reiterates the need to strictly follow standardized validation procedures. For a diagnostic model to be truly effective in a real-world medical setup, accuracy is not the only criterion for its success. It is imperative to incorporate the use of explainable artificial intelligence to make the models more interpretable and gain the confidence of medical professionals and regulatory bodies. It is also important to validate the models on external datasets to ensure that they generalize well to different patient populations in real-world settings. Future studies must focus on improving the robustness of models, making them more transparent in reporting results, and making experiments more reproducible. Standardized benchmarking tools and collaborative validation efforts will play a critical role in ensuring that machine learning breakthroughs are translated into safe and effective healthcare solutions.

REFERENCE

[1] Aggarwal R, Sounderajah V, Martin G, et al.
Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis.
npj Digital Medicine, 2021.
https://www.nature.com/articles/s41746-021-00438-z

[2] Liu X, Faes L, Kale AU, et al.
A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis.
The Lancet Digital Health, 2019.
https://www.sciencedirect.com/science/article/pii/S2589750019300232

[3] WU Q, GUO H, LI R, HAN J.
Deep learning and machine learning in CT-based COPD diagnosis: systematic review and meta-analysis.
International Journal of Medical Informatics, 2025.
https://www.sciencedirect.com/science/article/pii/S1386505625000150

[4] Takita H, Kabata D, Walston SL, et al.
Diagnostic performance comparison between generative AI and physicians: systematic review and meta-analysis.
npj Digital Medicine, 2025.
https://www.nature.com/articles/s41746-025-01102-5

[5] Hoodbhoy Z, Jiwani U, Sattar S, et al.
Diagnostic accuracy of machine learning models to identify congenital heart disease: a meta-analysis.
Frontiers in Artificial Intelligence, 2021.
https://www.frontiersin.org/articles/10.3389/frai.2021.708365

[6] Li WT, Ma J, Shende N, et al.
Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis.
BMC Medical Informatics and Decision Making, 2020.
https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01366-8

[7] Bečulić H, et al.
Sensitivity and specificity of ML and DL algorithms in thoracolumbar fracture diagnosis: systematic review and meta-analysis.
Brain and Spine, 2024.
https://www.sciencedirect.com/science/article/pii/S2772566924000123

[8] McGenity C, et al.
Artificial intelligence in digital pathology: diagnostic test accuracy systematic review and meta-analysis.
arXiv Preprint, 2023.
https://arxiv.org/abs/2301.01234

[9] Lu W, Tang X, Huang C, et al.
Diagnostic accuracy of ML and DL methods for detecting depression using speech features: meta-analysis.
BMC Psychiatry, 2025.
https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-025-04872-1

[10] Esteva A, Kuprel B, Novoa RA, et al.
Dermatologist-level classification of skin cancer with deep neural networks.
Nature, 2017.
https://www.nature.com/articles/nature21056

[11] Rajpurkar P, et al.
CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning.
arXiv, 2017.
https://arxiv.org/abs/1711.05225

[12] Ardila D, et al.
End-to-end lung cancer screening with deep learning.
Nature Medicine, 2019.
https://www.nature.com/articles/s41591-019-0447-x

[13] Topol EJ.
High-performance medicine: the convergence of human and artificial intelligence.
Nature Medicine, 2019.
https://www.nature.com/articles/s41591-018-0300-7

[14] Eriksen KG, et al.
Machine learning in cardiovascular disease diagnosis: systematic review.
European Heart Journal Digital Health, 2022.
https://academic.oup.com/ehjdh/article/3/1/25/6512223

[15] Kelly CJ, Karthikesalingam A, Suleyman M, et al.
Key challenges for delivering clinical impact with AI.
BMC Medicine, 2019.
https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1426-2

[16] Riley RD, et al.
Meta-analysis of diagnostic test accuracy studies.
BMJ, 2015.
https://www.bmj.com/content/351/bmj.h5527

[17] Higgins JPT, Thompson SG.
Quantifying heterogeneity in a meta-analysis (I^2 statistic).
Statistics in Medicine, 2002.
https://onlinelibrary.wiley.com/doi/10.1002/sim.1186

[18] Egger M, et al.
Bias in meta-analysis detected by a simple graphical test (Egger’s test).
BMJ, 1997.
https://www.bmj.com/content/315/7109/629

[19] Beam AL, Kohane IS.
Big data and machine learning in health care.
JAMA, 2018.
https://jamanetwork.com/journals/jama/article-abstract/2673158

[20] Gulshan V, et al.
Development and validation of a deep learning algorithm for diabetic retinopathy detection.
JAMA, 2016.
https://jamanetwork.com/journals/jama/fullarticle/2588763

[21] Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D.
Key challenges for delivering clinical impact with artificial intelligence.
BMC Medicine, 2019.
https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1426-2

[22] Sendak MP, D'Arcy J, Kashyap S, et al.
A path for translation of machine learning products into healthcare delivery.
EMJ Innovations, 2020.
https://www.emjreviews.com/innovations/article/a-path-for-translation-of-machine-learning-products-into-healthcare-delivery/

[23] Roberts M, et al.
Common pitfalls and recommendations for using machine learning to detect and diagnose COVID-19 using imaging data.
Nature Machine Intelligence, 2021.
https://www.nature.com/articles/s42256-021-00307-0

[24] Wynants L, et al.
Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal.
BMJ, 2020.
https://www.bmj.com/content/369/bmj.m1328

[25] Nagendran M, et al.
Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies.
BMJ, 2020.
https://www.bmj.com/content/368/bmj.m689