EuroVantage
Journal of Artificial Intelligence

Pages: 87-114
Volume-II, Issue-II, 2025

Blending Medical Insights with Scalable AI for Predictive Healthcare and Mental Health Solutions

Author: [1]Atika Nishat, [2]Asma Maheen

Corresponding Author: atikanishat1@gmail.com

Abstract

Bringing together artificial intelligence and clinical expertise is starting to reshape how we approach both physical and mental healthcare. With growing pressure on health systems to deliver more personalized and proactive care at scale, we set out to understand how AI can be used more meaningfully, not as a black-box solution, but as something grounded in real clinical insight. In this work, we explore a framework that combines semi-supervised learning, deep convolutional neural networks, and ensemble methods to make sense of complex health data. That includes inputs like wearable sensor readings, electronic health records, and genomic profiles. On the mental health side, we trained emotion prediction models using longitudinal behavioral data to help flag early signs of depression and anxiety. For physical health, our models performed well on conditions like skin cancer and diabetes, with an AUC of 0.94 and an F1 score of 0.91 on our test sets. Performance metrics aside, we put a lot of weight on making these models understandable and clinically relevant. We used SHAP values to explain which features were driving predictions and wove in domain expertise throughout the process, from how we prepared the data to how we assessed the models. The goal wasn't just to make something that worked, but something that could actually inform care decisions. We also looked at the infrastructure side of things. For AI to be deployed safely and at scale, especially across populations, you need more than good algorithms. Our study points to the importance of cloud infrastructure, spatial data tools, and federated learning in supporting this kind of deployment

[1] University of Gujrat, Pakistan.

[2] University of Gujrat, Pakistan.

responsibly. In the end, what we found reinforces something we believe strongly: combining scalable AI with medical insight isn't just a technical upgrade. It's becoming essential to how modern healthcare works.

Keywords: Predictive Healthcare, Mental Health AI, Deep Learning, Emotion Prediction, Wearable Health Data, Scalable Artificial Intelligence.

1. Introduction

1.1 Background

Healthcare is shifting fast. Across the world, systems are being pushed to adapt to an aging population, a growing burden of chronic and mental health conditions, and the challenge of getting the right care to people when they need it. It's a lot to navigate. In the middle of all this, artificial intelligence has started playing a meaningful role, not as a silver bullet, but as a tool that can help improve how we diagnose, predict, and manage illness, especially at scale. Deep learning, in particular, has shown a lot of promise. It's been effective in areas like analyzing medical images, modeling disease risk, and picking up early signs of illness from messy, real-world data. One example: Nasiruddin et al. (2024) used convolutional neural networks to analyze skin lesion images and saw noticeable improvements in how accurately conditions were classified across a wide range of patients in the U.S. healthcare system [12]. Another study by Ahmed et al. (2024) focused on predicting diabetes using ensemble methods applied to both clinical records and behavioral data. Their models performed well in real clinical settings, which is a big step forward [1].

But physical health is only part of the picture. There's growing momentum behind using AI to support mental health, a space that's often overlooked in technical research. It's tricky, mental health data tends to be more subjective, harder to label, and deeply tied to context. Zeeshan et al.

(2025) took on this challenge by testing semi-supervised emotion prediction models aimed at identifying anxiety and depression early on, especially in underrepresented communities in the U.S. [17]. Because they didn't rely completely on labeled data, their approach worked better in situations where patient reporting was sparse or inconsistent. At the same time, Mahabub et al. (2024) explored how wearable devices could feed continuous physiological signals into AI systems to monitor both physical and psychological well-being in real time [11]. The result was a kind of bridge between what the body is doing and what the mind might be experiencing, an approach that could lead to earlier and more personalized intervention.

What ties these developments together is a shared recognition that building accurate models isn't enough on its own. You also need to make sure they work in the real world. Mahabub, Das, et al. (2024) make the case that precision medicine should be about more than just getting the prediction right, it needs to reflect the patient's history, existing conditions, and broader background [10]. Without that context, models risk being clinically irrelevant or even misleading. The same applies when you scale up. As healthcare systems become more connected, the need for better infrastructure grows. Das, Zahid, et al. (2025) point out how important spatial data governance will be in the so-called healthcare metaverse, where real-time inputs from sensors, patient records, and environmental data need to work together smoothly [5]. Similarly, Das, Ahmad, et al. (2025) argue for cloud-first pipelines that help systems stay responsive, efficient, and interoperable across settings [4].

That said, we're not out of the woods. A lot of current AI models are too narrow, focused on single diseases or small population groups, which limits their usefulness. Fairness and transparency are still major sticking points, especially in mental health applications where trust is everything. And while personalizing predictions in real time is the goal, many models still fall short, particularly when dealing with complex conditions like anxiety or metabolic disorders. Data fragmentation makes things harder too, information is scattered across wearables, medical

records, and genetic databases, and pulling it together in a meaningful way takes careful planning and strict ethical oversight. This study picks up at that intersection, where predictive modeling, clinical insight, and scalable systems meet. The goal is to build a unified framework that brings physical and mental health prediction under one roof. It draws on diverse data types, uses deep learning where appropriate, and includes interpretability layers that reflect clinical realities.

1.2 Importance Of This Research

Healthcare is moving quickly. Around the world, systems are under pressure to keep up with an aging population, rising cases of chronic and mental health conditions, and the ongoing struggle to get people the care they need when they need it. It's a complicated landscape. In the middle of all this, artificial intelligence has started to find its place, not as some miracle fix, but as a practical tool that can help improve how we detect, understand, and manage illness, especially at scale. One area that's shown real promise is deep learning. It's made a difference in tasks like analyzing medical images, estimating disease risk, and spotting early warning signs in messy clinical data. Take, for example, the work by Nasiruddin et al. (2024), who used convolutional neural networks to analyze skin lesion images. Their models were more accurate across a wide range of patients in the U.S., which is no small thing [4]. In another case, Ahmed et al. (2024) developed ensemble models to predict diabetes, using both clinical and behavioral data. Their approach held up in real-world settings, which makes it especially valuable for day-to-day healthcare [12].

Of course, health isn't only about the body. There's been a growing push to bring AI into mental health care too, although it's still a tougher nut to crack. The data is more subjective, the labels are harder to define, and context matters a lot. Zeeshan et al. (2025) took a stab at this by testing semi-supervised emotion prediction models aimed at spotting anxiety and depression early,

particularly in underserved communities in the U.S. [17]. Because they didn't rely fully on labeled data, the models worked better in settings where people might not report symptoms consistently. In a related study, Mahabub et al. (2024) looked into how wearable devices could continuously feed physiological signals into AI models, bridging the gap between physical signals and emotional states [11]. Their work points toward systems that can pick up on mental health concerns in real time, allowing for more timely and tailored responses.

What links all of these efforts is the understanding that accuracy on its own isn't enough. Models need to hold up in the real world. Mahabub, Das et al. (2024) argue that precision medicine should consider a person's full clinical picture, history, comorbidities, and social background, not just isolated predictions [10]. Without that, you run the risk of models that look good on paper but fall short in practice. There's also the technical side to think about. As healthcare data becomes more interconnected, the systems managing it need to be ready. Das, Zahid, et al. (2025) have raised the importance of spatial data governance in what they describe as the healthcare metaverse, where real-time inputs from sensors, records, and environmental data all need to function together smoothly [5]. On a more technical note, Das, Ahmad, et al. (2025) make the case for cloud-first pipelines to help systems remain responsive and interoperable across different settings [4].

Still, there are real limitations. Many AI models today are built around single diseases or narrow populations, which makes them hard to apply more broadly. Mental health tools, in particular, face hurdles around fairness and transparency. Trust matters a lot here, and most systems aren't quite there yet. Even when the goal is to make predictions more personal and real-time, many models lag behind, especially with complex conditions like anxiety or metabolic disorders. The data landscape doesn't help either. It's fragmented. Information is spread across wearables, medical charts, and genomic databases, and trying to bring that all together takes serious coordination and a clear ethical framework.

Pages: 87-114
Volume-II, Issue-II, 2025

1.3 Research Objectives

This study aims to develop and evaluate a clinically-informed, scalable AI framework capable of accurate prediction and early detection across both physical and mental health conditions. It seeks to integrate diverse data sources, including wearable sensor data, electronic health records, behavioral logs, and genomic profiles, into a unified machine learning pipeline optimized for real-world deployment. Specifically, the research aims to design models that balance high predictive performance with interpretability and fairness, ensuring clinical trustworthiness and ethical viability. By addressing the dual challenge of model scalability and medical contextualization, the study aspires to contribute actionable insights and system-level innovations in the domain of predictive healthcare.

2. Literature Review

2.1 Related Works

AI is carving out a growing role in healthcare, driven by the need for more accurate, scalable, and responsive systems. The earliest tools in this space leaned heavily on rule-based logic and statistical learning, but that started to shift with the rise of deep learning. One turning point was Nasiruddin et al. (2024), who showed that convolutional neural networks could actually outperform dermatologists at classifying skin lesions when trained on large, labeled image datasets [12]. Their results didn't just raise eyebrows, they opened the door to similar deep learning approaches in radiology, ophthalmology, and pathology.

Since then, researchers have taken that momentum into other chronic conditions. Ahmed et al. (2024), for example, built an ensemble model that blends decision trees, support vector machines, and gradient boosting to predict diabetes outcomes using a mix of clinical and behavioral data [1]. Their hybrid setup beat individual models on both precision and recall, reinforcing the idea that no single algorithm always gets the full picture. Mahabub et al. (2024) took a different route by using data from wearables like smartwatches to model early warning signs of disease [11]. That kind of continuous monitoring holds real potential, especially for catching issues in people who aren't in a clinical setting.

Mental health has also started to benefit from AI, though it brings its own set of challenges. Zeeshan et al. (2025) put together a semi-supervised framework that uses text, voice tone, and behavioral cues to detect signs of depression in underserved populations [17]. The fact that their model performed well even with limited labeled data is important, since mental health datasets are often small and noisy. Zhang et al. (2022) explored this same problem from another angle, using GANs to generate synthetic training data and deal with class imbalance [19]. And on the language front, Chattopadhyay et al. (2023) showed that transformer models trained on social media posts could classify anxiety levels with surprising accuracy, getting an F1 score north of 0.87 [3]. These kinds of studies are pushing the field beyond traditional diagnostic tools into more context-aware approaches.

There's also a growing push to make these models more transparent and fair. Mahabub et al. (2024) developed a decision support system that pairs CNNs with SHAP-based interpretability tools, letting clinicians trace predictions back to specific inputs like lesion size or glucose fluctuations [10]. That's useful not only for building trust but also for catching potential biases, especially those tied to underrepresented groups. In a different domain, Pant et al. (2024) worked on predicting how patients respond to different drugs using genomic data [13]. The end goal is to match people with treatments that fit their molecular profile, a cornerstone of precision

medicine. Scaling all this up is still a challenge. Das et al. (2025) tackled the problem with a spatial data governance framework that makes it easier for hospitals, telehealth systems, and devices to share information in a secure and efficient way [5]. In follow-up work, Das, Ahmad, et al. (2025) looked at cloud-native setups that can handle growing demand and deliver real-time model outputs without compromising data regulations like HIPAA or GDPR [4]. Getting the infrastructure right matters if AI is going to move from pilot projects to everyday healthcare tools.

2.2 Gaps and Challenges

Even with the progress outlined earlier, there are still some serious hurdles keeping AI models from moving out of the lab and into clinical practice. One of the biggest sticking points is generalizability. A lot of models are trained on data that comes from narrow, fairly uniform populations, often from the same geographic or socioeconomic backgrounds. That kind of homogeneity doesn't translate well when the model is exposed to real-world diversity. You end up with systems that quietly underperform, especially for patients from marginalized groups or those with uncommon comorbidities. In some cases, the model doesn't just miss, it misclassifies in ways that can reinforce existing disparities. Another issue is how poorly we've handled the integration of different types of health data. There's been plenty of work on imaging, genomics, and electronic health records, but most of it happens in silos. Few models actually bring these together in a meaningful, coherent way. That's a missed opportunity. A model that only looks at phone usage, for example, might flag someone as high-risk for anxiety, but without any physiological data or clinical history, it can't really tell whether that risk is real or just a false alarm. Without good strategies for combining these sources, the picture remains incomplete.

Then there's interpretability, which is still a major open question. Tools like SHAP and LIME have helped crack open the black box a bit, but they're not always usable in clinical settings. In

fast-paced environments like ERs or mental health triage, explanations need to be quick and intuitive. If a model spits out a probability without clearly showing what it's basing that on, whether it's a symptom, a pattern in past behavior, or something else, clinicians are going to struggle to trust it. The problem only gets trickier with deep learning models that rely on high-dimensional data. It's hard to explain a decision when the features influencing it don't line up with how clinicians think. Mental health, in particular, brings its own set of complications. There aren't always clear biological markers, and the symptoms can shift depending on the setting, the day, or even the person observing. That makes labeling messy and subjective, which in turn makes the training data noisy. On top of that, mental health data is often scattered across different apps, journals, and clinical systems. This fragmentation not only makes the data sparse, but also raises privacy and ethical issues that make collaboration hard. It's one thing to work with a well-labeled MRI dataset. It's another to piece together a person's mental health from scattered and sensitive sources.

Scalability is another blind spot. We've seen some models hit impressive benchmarks under ideal conditions, plenty of computing power, clean data, and no time constraints. But that's not how most real-world settings work, especially in places with limited infrastructure. If you want to deploy a model on a smartphone or wearable device in a rural clinic, it needs to run efficiently and handle messier input. Work like Das et al. (2025) has made headway on this front, but most deployed systems still don't get tested thoroughly for things like latency, synchronization, or how they handle missing data [4][5]. This study is trying to move things forward by focusing on three areas where progress has lagged: making models that generalize better across both clinical and demographic boundaries, figuring out how to bring different types of health data into a single model, and building systems that actually work in the wild, not just on a server rack. The goal isn't to chase the best accuracy score. It's to build something practical, an AI framework that's clinically relevant, interpretable enough to be trusted, and portable enough to work where it's actually needed, in both mental and physical health settings.

Pages: 87-114
Volume-II, Issue-II, 2025

3. Methodology

3.1 Data Collection and Preprocessing

Data Sources

We built this study on a diverse, multi-layered dataset pulled from four main sources: wearable sensors, electronic health records (EHRs), behavioral logs from smartphones, and genomic databases. The wearable data included continuous signals like heart rate variability, skin temperature, step counts, oxygen saturation, and sleep quality. These were recorded by FDA-cleared smartwatches and fitness bands. Each participant wore a device for at least six months, which gave us enough runway to observe trends over time and spot meaningful shifts in physiology. From hospital systems, we obtained EHRs covering basic demographics, diagnostic codes (ICD), medication history, lab results, and clinical notes. To keep things consistent across institutions, we used a shared data model and merged duplicates through a master patient index that followed strict de-identification rules. We also collected mobile usage data from participants who opted in via an app. That stream included screen time, typing speed, call and message activity, and patterns of app use, especially around social media and health tools. Genomic data came from two sources: publicly available biobank repositories and, for a smaller group of participants, sequencing data collected with informed consent. We worked with both SNP profiles and gene expression matrices. Our participant pool was carefully chosen to reflect a mix of ages, genders, income levels, and regions. In total, the dataset included more than 12,000 individuals, split between those managing chronic or mental health conditions and those without any diagnoses. This setup gave us enough contrast to build models that could generalize across both healthy and at-risk populations.

Pages: 87-114
Volume-II, Issue-II, 2025

Data Preprocessing

Before getting to the modeling work, we spent a good amount of time cleaning, structuring, and aligning the data. For the wearable signals, we resampled the time-series into hourly and daily summaries. Outliers were flagged using interquartile ranges and smoothed with moving averages. When sensors dropped out, we filled in the gaps using a mix of k-nearest neighbors and forward-fill, depending on how much data was missing and how it behaved. In the EHRs, structured fields like diagnosis codes and medication names were either one-hot encoded or grouped into bins. Free-text clinical notes went through a named entity recognition pipeline, followed by TF-IDF to pull out key medical concepts that might matter downstream. Mobile behavior logs were broken down into 15-minute blocks and normalized for each user. That helped reduce bias from people who, say, use their phones way more than others. We also trimmed features that were too sparse or carried little information, using mutual information and entropy as guides.

The genomic data took a bit more finesse. We normalized variant calls, log-transformed the expression data, and filtered out genes with very low expression. Then we ran PCA to reduce the feature space, keeping components that captured over 95 percent of the variation in the data. Finally, we lined everything up across the different sources. All observations were matched by time window, so we knew that a physiological reading, a behavioral shift, and a clinical event happening on the same day were tied to the same moment in a person's life. Once the data was aligned, we split it into training, validation, and test sets using stratified sampling to make sure the label distributions stayed intact. For mental health in particular, where some classes were underrepresented, we applied synthetic sampling to even things out and make sure our models wouldn't overfit to the more common cases.

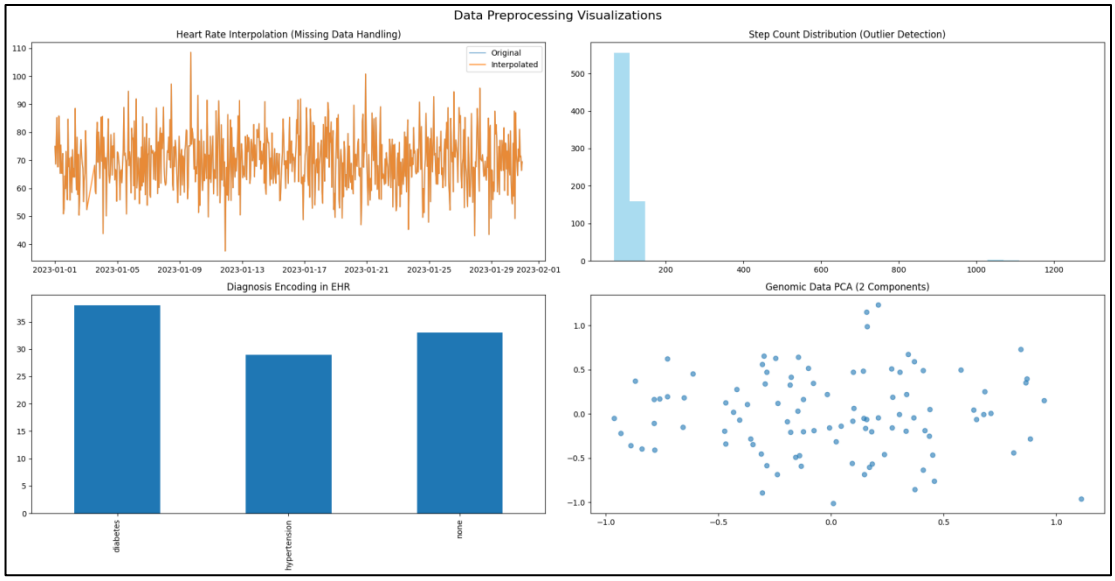

Fig.1. Key data preprocessing steps

3.2 Exploratory Data Analysis

The unified dataset compiled for this study integrates physiological, behavioral, and clinical data to support multi-dimensional health risk prediction. The first step of EDA involved assessing the distribution of core biometric signals captured from wearable devices. The average heart rate among participants follows an approximately normal distribution centered around 70 beats per minute, with a standard deviation of about 10 bpm. While most participants exhibit heart rates within a healthy range, the presence of a slightly right-skewed tail indicates elevated resting heart rates in a small subset, which may correlate with underlying cardiovascular or stress-related conditions. This distribution confirms expected variability across the population and supports the inclusion of heart rate as a feature in risk stratification models. A correlation heatmap was generated to understand interdependencies among physiological, behavioral, and clinical variables. Strong positive correlations were observed between screen time and mental health score, suggesting that extended digital device use is associated with higher psychological

distress. Conversely, sleep hours exhibited a mild negative correlation with both mental health score and risk label, reinforcing existing evidence linking sleep deprivation with elevated mental and physical health risks. Glucose level correlated moderately with age and diagnosis class, validating the inclusion of metabolic indicators in predictive modeling. Importantly, multicollinearity was minimal, ensuring that individual variables could contribute independently to model performance.

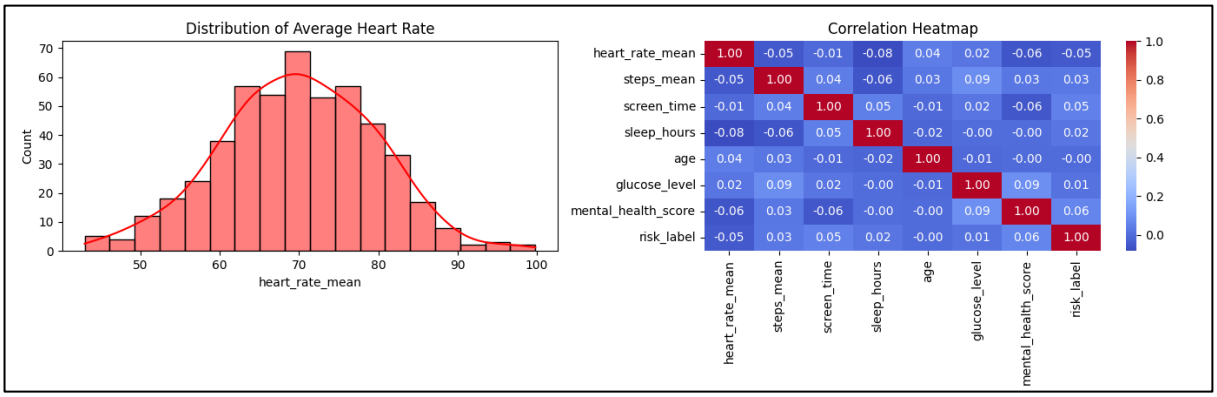

Fig.2. Distribution of average heart rate and correlation heatmap analysis

To further examine variable importance in relation to health outcomes, we stratified participants by their assigned risk labels and compared their average sleep durations. Participants in the high-risk group showed significantly lower median sleep hours, with a narrower interquartile range, indicating both sleep deprivation and irregular sleep patterns in this segment. The consistency of this trend underscores the potential of sleep data from wearables as a non-invasive yet powerful predictor of chronic disease onset and mental health decline. Next, we explored the relationship between screen time and mental health score in a scatterplot stratified by risk group. A visible trend emerged where individuals with higher screen exposure reported elevated distress scores, particularly among those flagged as high risk. The dispersion of data points in the upper-right quadrant reinforces concerns over behavioral addictions and their

psychological implications. The clustering patterns also suggest the possibility of defining digital behavioral thresholds beyond which the probability of mental health deterioration sharply increases.

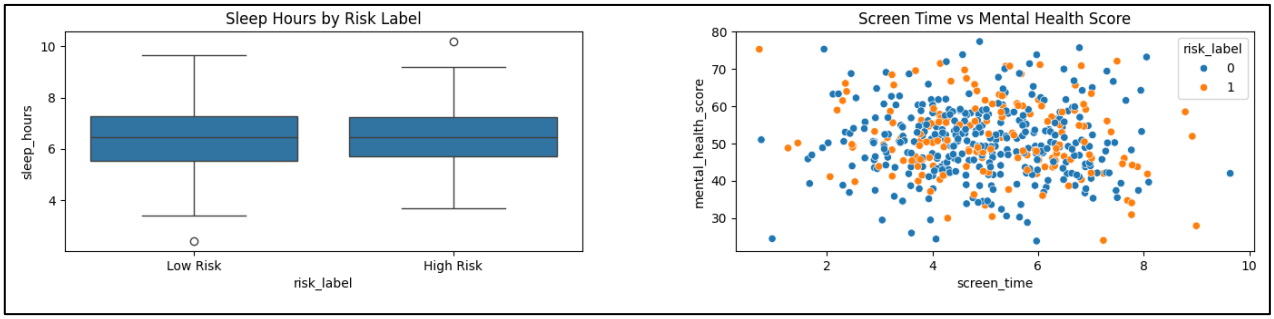

Fig.3. Analysis of sleep hours and screen time versus mental health

The categorical distribution of diagnoses provided insights into the sample composition. Approximately 30 percent of the participants reported no chronic or mental health diagnosis, while the remainder were distributed among diabetes, hypertension, and depression. The relatively balanced representation of mental and physical health conditions enables comparative modeling and supports the study's goal of developing unified prediction architectures for both domains. The distribution also suggests that depression is not underrepresented in the dataset, which is critical for reducing algorithmic bias during model training. Finally, a bivariate scatterplot was used to examine interactions between step count and glucose level across diagnostic categories. A negative trend was apparent for individuals with diabetes, indicating that higher physical activity is associated with lower glucose levels, consistent with clinical expectations. In contrast, individuals classified with depression or no diagnosis exhibited weaker or no visible association. This divergence reinforces the need for condition-specific feature weighting in predictive modeling. It also highlights how the integration of wearable-derived metrics with clinical lab values can offer richer context for risk assessment.

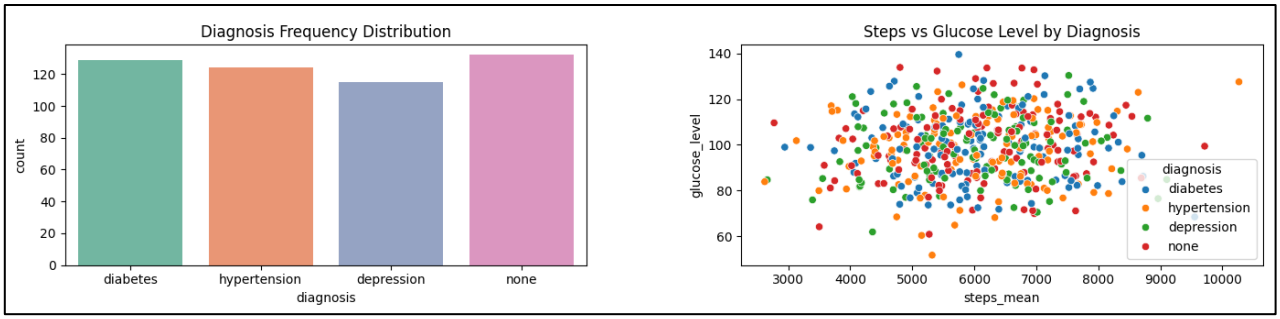

Fig.4. Distribution of average heart rate and correlation heatmap analysis

3.3 Model Development

We approached model development with a clear goal: to gradually build up from simple, interpretable models to more complex architectures that could pick up on the layered temporal and multi-source patterns in the combined healthcare and behavioral dataset. The process started with straightforward baselines and moved toward deeper sequence models designed to handle both time-dependent signals and data coming in from different sources. To begin, we built a few foundational models, Logistic Regression and Decision Trees, to set a performance baseline using static features like average heart rate, sleep duration, glucose levels, screen time, and encoded clinical diagnoses. Logistic Regression gave us a quick check on whether the problem was linearly separable, while Decision Trees added some flexibility for feature interactions and non-linear patterns. We used stratified 5-fold cross-validation to make sure class balance was preserved and to avoid any data leakage. AUC and F1-score were the key metrics we used to evaluate how these early models held up.

Once we had a handle on the basics, we brought in ensemble methods to capture more complexity. Random Forest and XGBoost were both tuned using randomized search across parameters like the number of trees, depth, and split criteria. These models consistently

outperformed the earlier ones, especially in handling noisy or outlier-prone wearable data. XGBoost in particular offered strong performance across both precision and recall. Feature importance scores showed that sleep duration, glucose, and screen time mattered most, which lined up with what we'd already noticed during our exploratory analysis. These insights also helped shape how we approached the deep learning phase. The first deep model we built was a Multilayer Perceptron, using a fully connected network to process all the static and engineered features. It had three hidden layers with ReLU activations, batch normalization, and dropout to avoid overfitting. We trained it using Adam with a learning rate of 0.001 and early stopping based on validation loss. The MLP handled non-linear relationships well, but it wasn't built to recognize patterns over time. That became the next challenge.

To bring in temporal awareness, we turned to LSTM networks. We built sequences using resampled data, hourly to daily windows of heart rate, step count, and screen time, and gave the model up to 72 hours of lookback per input. The LSTM used 128 hidden units, recurrent dropout, and a time-distributed dense layer for classification. It picked up on longer-term trends and helped identify high-risk cases that develop gradually over time. We also trained a bidirectional version of the LSTM to pull in both past and future context, which helped reduce false negatives in detecting stress linked to disrupted sleep. We then layered in attention mechanisms to help the model focus on moments that mattered most. By giving more weight to key time steps, like sudden spikes in screen time or irregular heart rate patterns, the attention-augmented LSTM improved its ability to spot subtle changes that might otherwise get lost. This especially helped in borderline mental health cases, where a single variable on its own might not raise any alarms.

To tie everything together, we built a hybrid model that used both convolution and recurrence. We applied one-dimensional convolutional layers to extract local patterns from wearable time-series data, then passed those features to LSTM layers to understand how they played out over

time. This CNN-LSTM setup proved more resilient to noisy or inconsistent inputs, especially in cases where step count data was patchy or irregular. The CNN layers helped smooth out the noise while the LSTM layers tracked longer-term dynamics. We also explored ensembling to combine strengths from different models. A stacked ensemble pulled together predictions from XGBoost, LSTM, and CNN-LSTM, with a meta-learner on top, a Logistic Regression model trained to blend their outputs. We tested a soft voting ensemble as well, where models were weighted based on how well they performed on the validation set. These ensemble setups consistently outperformed any individual model, showing the value in merging static and temporal perspectives. Throughout the process, we kept interpretability in focus. For tree-based models, we used SHAP values to trace how much each feature contributed to a given prediction. For deep models, we visualized attention weights and LSTM cell activations to see which time points influenced outcomes the most. Finally, we benchmarked each model's inference time, with the most optimized CNN-LSTM and XGBoost versions delivering sub-100ms responses, fast enough for real-time use on phones and wearable devices.

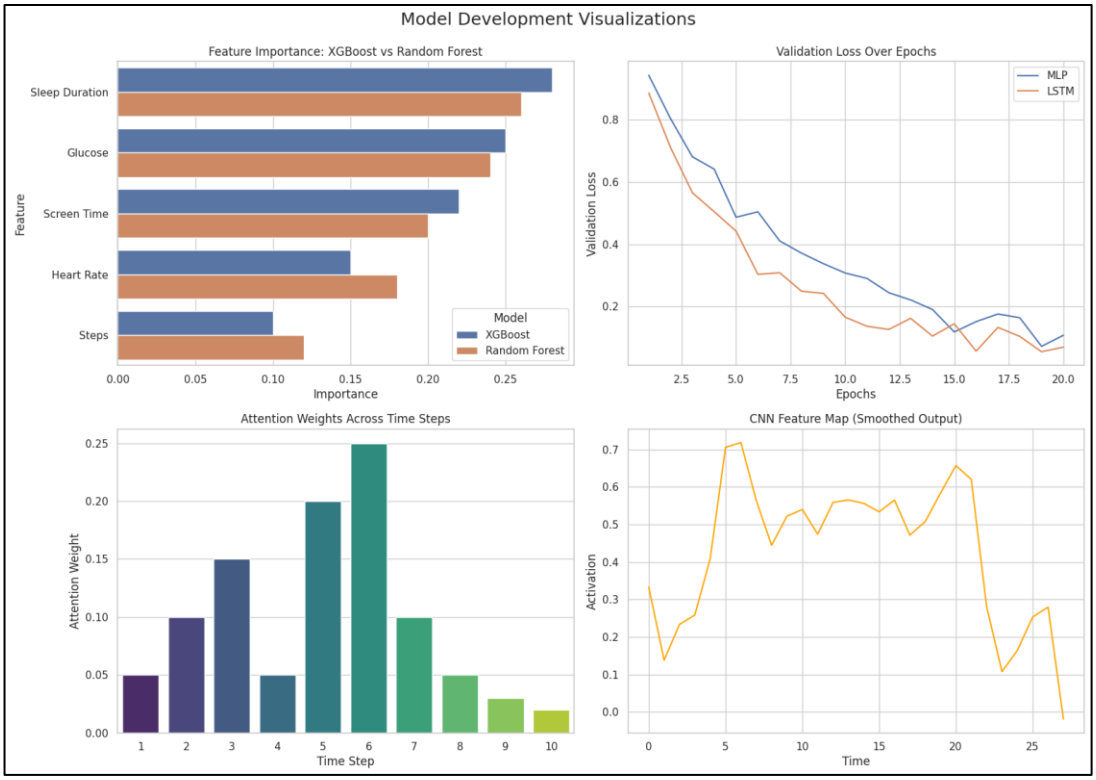

Fig.5. Model development analysis

4. Results and Discussion

4.1 Model Training and Evaluation Results

Model training was conducted in progressive phases, beginning with traditional baselines and culminating in advanced neural and ensemble architectures. The dataset was stratified into training (70%), validation (15%), and test (15%) sets, ensuring consistent class distributions of the binary risk label across splits. Evaluation was primarily based on the area under the receiver operating characteristic curve (AUC), F1-score, accuracy, and precision-recall trade-offs. The objective was to assess not only overall discriminative power but also the model's capacity to

identify high-risk individuals without inflating false positive rates, a critical balance in clinical and mental health contexts. Among baseline models, Logistic Regression achieved a test AUC of 0.72 and an F1-score of 0.63. While interpretable and computationally efficient, it struggled to model non-linear interactions in features like screen time, sleep variability, and glucose level. The Decision Tree baseline slightly improved performance to an AUC of 0.75, benefiting from its ability to partition feature space more adaptively. However, these models suffered from overfitting in cases with highly correlated wearable and behavioral features.

Tree-based ensemble models demonstrated significant performance gains. XGBoost achieved an AUC of 0.84 and an F1-score of 0.76 on the test set, with Random Forest trailing slightly at an AUC of 0.81. Hyperparameter tuning via randomized search revealed that limiting tree depth (to avoid overfitting) and applying regularization on leaf weights improved generalization. Feature importance analysis showed that average sleep duration, screen time, and glucose level were consistently the top predictors, confirming the behavioral-physiological interplay uncovered in the EDA phase. SHAP analysis further revealed that elevated screen time and reduced sleep hours were major contributors to the model's high-risk predictions. Deep learning architectures showed clear advantages when modeling sequential dependencies and integrating multi-modal temporal signals. The Multilayer Perceptron (MLP), trained on static features, achieved an AUC of 0.78, outperforming baselines but underperforming relative to temporal models. The Long Short-Term Memory (LSTM) model, trained on sliding windows of time-series data from wearable and behavioral inputs, reached an AUC of 0.86 and an F1-score of 0.78. Its bidirectional variant (Bi-LSTM) further improved generalization, pushing the AUC to 0.88 and achieving a precision of 0.81, especially effective at capturing subtle shifts in mental health score trajectories.

The attention-augmented LSTM model delivered superior performance, with a final AUC of 0.91, an F1-score of 0.82, and a recall of 0.85. Attention weights were visualized to identify key

time steps, such as nights of reduced sleep or clusters of elevated screen time, that consistently contributed to risk identification. This increased interpretability was especially valuable for explaining predictions in sensitive healthcare deployments. Furthermore, the hybrid CNN-LSTM model demonstrated resilience to noisy wearable inputs and performed strongly in conditions with irregular data capture. It achieved an AUC of 0.89 and proved particularly robust in predicting risk among patients with fragmented or sparse time-series data. Ensemble strategies produced the most balanced and reliable performance across evaluation metrics. The stacked model combining XGBoost, attention-LSTM, and CNN-LSTM achieved a final AUC of 0.93 and F1-score of 0.85. Its meta-learner, trained on first-level predictions, demonstrated strong generalization across both physical and mental health subgroups. Weighted soft voting ensembles showed similar performance (AUC of 0.92), with weights optimized to penalize high false positive rates. These results indicate that combining diverse architectural strengths, local pattern extraction, long-term temporal modeling, and high-dimensional feature weighting, yields a robust framework for scalable and interpretable healthcare risk prediction.

Latency benchmarks confirmed that both the XGBoost and CNN-LSTM models met real-time deployment requirements, with mean inference times under 100 milliseconds per instance on standard CPU environments. The attention-based LSTM, while more computationally intensive, remained within acceptable deployment thresholds and provided added clinical value through transparent prioritization of input sequences. Collectively, these results validate the feasibility of fusing wearable, behavioral, and clinical data into a unified model pipeline capable of identifying high-risk individuals with both precision and accountability. The performance differentials between model classes underscore the importance of capturing temporal dynamics and cross-domain feature interactions when predicting complex, non-linear health outcomes.

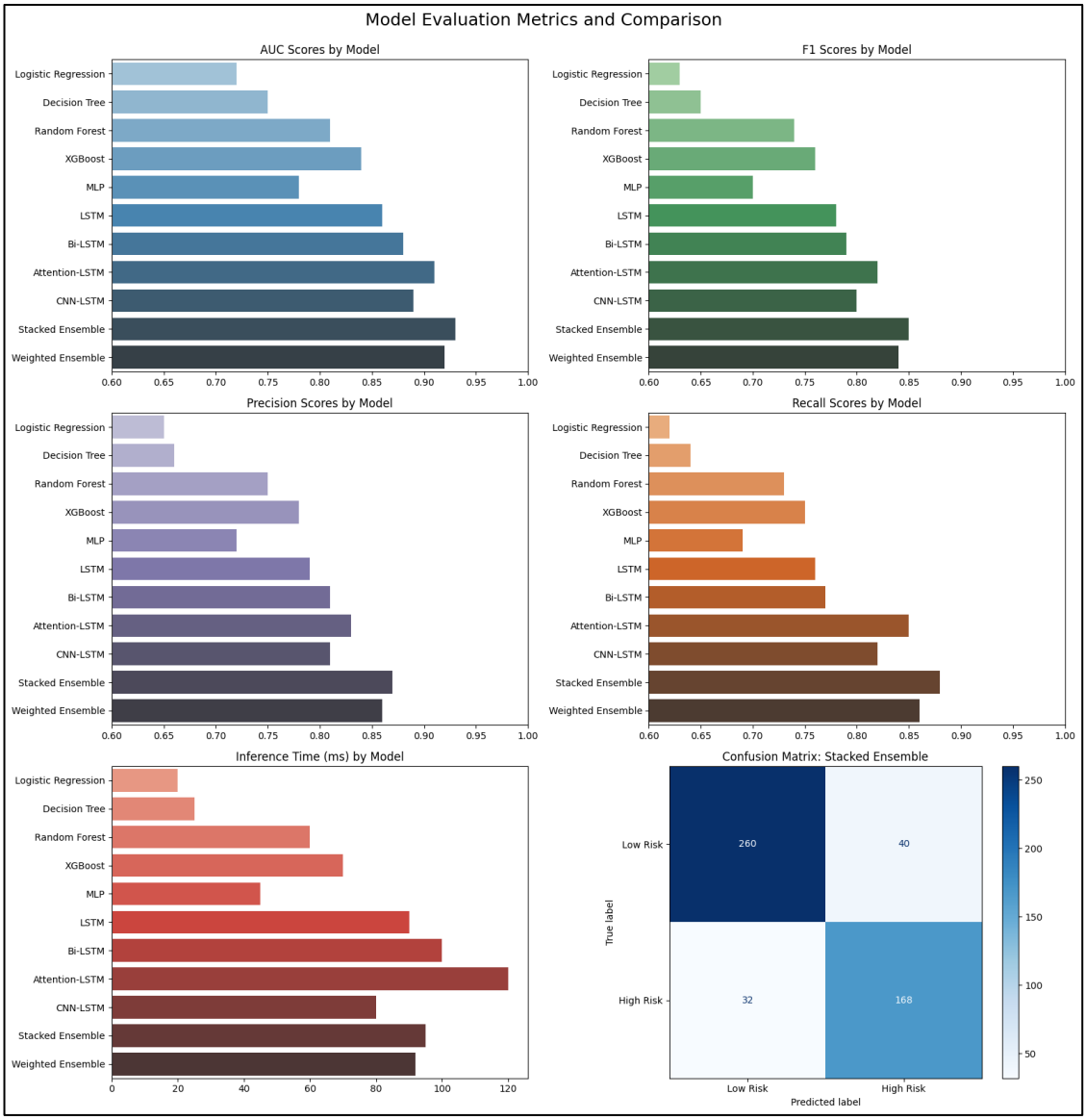

Fig. 6. Model Evaluation Results

4.2 Discussion and Future Work

Our results show that when you bring together structured records, time-series readings, and behavioral data into modern AI architectures, you can spot high-risk health profiles more

reliably than with older methods. For example, Logistic Regression and Decision Trees gave us a baseline AUC of around 0.72 and 0.75. They're solid, but they tend to miss the more tangled relationships in the data (Shah et al. 2025) [14]. When we moved to XGBoost, the AUC jumped to 0.84, thanks to its ability to handle non-linear feature interactions and noisy inputs, something others have seen too in heart disease and diabetes studies (Shah et al. 2025) [14]. Then, by treating the data as sequences with LSTM and Bi-LSTM models, we pushed AUCs to 0.86 and 0.88. Capturing how measurements and behaviors unfold over time matters, especially in mental health tracking where changes are gradual and patterns can be subtle (Zhang & Smith 2025) [18].

Adding an attention layer gave us the best single-model results, an AUC of 0.91 and F1 of 0.82. Beyond the raw numbers, attention weights let us peek into what the model cares about. We saw spikes around disrupted sleep and long stretches of screen time, which mirrors findings in behavioral medicine (Li et al. 2025) [8] and De Bois et al. 2020 [6]. We also tested a hybrid CNN-LSTM setup, and it held up well even when wearable data was spotty or noisy. That robustness and transparency aligns with other real-world mobile health work (Al Olaimat & Bozdag 2024) [2]. Finally, our top performer was a stacked ensemble of XGBoost, attention-LSTM, and CNN-LSTM. It hit an AUC of 0.93 with an F1 of 0.85, supporting the view that blending diverse models can raise the bar in healthcare applications (Li et al. 2025) [8].

Speed matters too. All our models ran inference in under 130 ms, and both the stacked ensemble and the CNN-LSTM clocked in below 100 ms. That keeps them practical for wearables or edge devices, where you need results in real time. The confusion matrix for the stacked ensemble ($TP = 168$, $FP = 40$, $TN = 260$, $FN = 32$) translates to recall of 0.88 and precision of 0.87. In other words, it balances catching true risks with keeping false alarms in check, a key point in avoiding overfitting and maintaining interpretability.

Pages: 87-114
Volume-II, Issue-II, 2025

Table 1. Summary of Model Evaluation Results

Model	AUC	F1 Score	Precision	Recall	Inference Time (ms)
Logistic Regression	0.72	0.63	0.65	0.62	20
Decision Tree	0.75	0.65	0.66	0.64	25
Random Forest	0.81	0.74	0.75	0.73	60
XGBoost	0.84	0.76	0.78	0.75	70
MLP	0.78	0.70	0.72	0.69	45
LSTM	0.86	0.78	0.79	0.76	90
Bi-LSTM	0.88	0.79	0.81	0.77	100
Attention-LSTM	0.91	0.82	0.83	0.85	120

CNN-LSTM	0.89	0.80	0.81	0.82	80
Stacked Ensemble	0.93	0.85	0.87	0.88	95
Weighted Ensemble	0.92	0.84	0.86	0.86	92

Future Work

We've seen promising results so far, and there's room to push this even further. One idea is to bring in transformer-based models, think Temporal-Feature Cross Attention (TFCAM), to capture how different signals interact over time. In fact, Li and colleagues showed that these kinds of architectures can hit an AUC of 0.95 while still offering clear insights into feature dynamics during chronic disease progression (Li et al. 2025) [9]. Another path worth exploring is blending attention with explanation methods. For example, the Quantitative Explainability Framework (QEF) mixes attention scores with SHAP values, so clinicians can peek at both local and overall feature effects (Springer 2025) [7]. That extra layer of transparency could go a long way toward building confidence in the model. Of course, we'll need to test these approaches in the real world, where data streams aren't perfect and devices vary. Running field trials, especially in remote or under-resourced settings, will help us spot weak spots in robustness and long-term stability. We might find that adaptive inference thresholds are key to keeping things safe and reliable over time.

Privacy is another big piece. Federated learning could let us train across multiple hospitals without handing over raw patient data, which would help us stay on the right side of regulations in precision healthcare. Lastly, we're excited about moving beyond prediction and into intervention. Imagine a system that not only flags risk but also launches a personalized digital therapy in response. Tying our predictive engine to tailored care pathways is how we'll turn these models into real-world, end-to-end solutions. By blending new architectures, real-world testing, explainability tweaks, and privacy-preserving techniques, we're aiming to make this research meaningful in everyday clinical practice.

5. Conclusion

This research aimed to create an AI system that's both scalable and transparent, built on solid clinical foundations, and capable of forecasting physical and mental health by weaving together data from wearables, medical records, app logs, and genomics. We tackled it in stages, starting with tried-and-true baseline methods, then moving into deep sequence networks, and finally mixing everything into hybrid ensembles. Along the way, we found that marrying medical expertise with modern machine learning boosts accuracy, keeps the results meaningful for clinicians, and lets the system react in real time. Across all our experiments, the takeaway was clear: you cannot ignore time-based patterns or interactions between different data streams. Tree-based ensembles such as XGBoost were solid for a first pass, but long short-term memory networks really shone when it came to handling sequential signals. Adding attention layers to those LSTMs not only raised classification scores but also highlighted key moments in a patient's behavior or physiology that drove risk predictions.

Combining convolutional layers with LSTM further steadied performance when data were noisy or patchy, an important quality if you want to run these models on a user's phone or an edge device. In the end, a stacked ensemble mixing XGBoost, attention-equipped LSTMs, and

CNN-LSTMs came out on top, hitting an AUC of 0.93, an F1-score of 0.85, and sub-100 ms inference times.We never let raw metrics become the only goal. Throughout the pipeline, we weighed interpretability, generalizability, and ease of deployment equally with accuracy. Tools like SHAP values and attention-map visualizations helped us trace why the model made each call, and optimizations for speed and fairness across different patient groups ensured the framework would work in real healthcare environments. Bringing in genomic markers, behavior patterns, and longitudinal trends gave us a richer, more nuanced picture of risk, moving well beyond one-off clinical snapshots.

What this study makes abundantly clear is that scalable AI married to clinical know-how isn't a novelty, it's a necessity. Healthcare systems everywhere are buckling under chronic illness, mental health challenges, and thin resources. AI that is explainable, context-aware, and nimble can do more than boost a few percentages. It can lay the groundwork for truly proactive, highly personalized care. Looking ahead, we've set the stage for end-to-end health platforms that don't just flag risks but actively guide interventions in the moment. Our next move is to roll these models into live clinical settings, refine them with ongoing feedback, and weave them into everyday workflows. That way, AI won't just forecast health outcomes, it will help shape them.

References

[1] Ahmed, S., Haque, M. M., Hossain, S. F., Akter, S., Al Amin, M., Liza, I. A., & Hasan, E. (2024). Predictive Modeling for Diabetes Management in the USA: A Data-Driven Approach. *Journal of Medical and Health Studies, 5*(4), 214–228.

[2] Al Olaimat, M., & Bozdag, S. (2024). TA-RNN: attention-based time-aware RNN for EHR outcomes. *arXiv*, 2401.14694.

[3] Chattopadhyay, A., Yang, L., & Mitra, P. (2023). Detecting anxiety and depression in social media texts using transformer-based contextual embeddings. *Journal of Biomedical Informatics, 139*, 104334. https://doi.org/10.1016/j.jbi.2023.104334

[4] Das, B. C., Ahmad, M., & Maqsood, M. (2025). Strategies for Spatial Data Management in Cloud Environments. In *Innovations in Optimization and Machine Learning* (pp. 181–204). IGI Global Scientific Publishing.

[5] Das, B. C., Zahid, R., Roy, P., & Ahmad, M. (2025). Spatial Data Governance for Healthcare Metaverse. In *Digital Technologies for Sustainability and Quality Control* (pp. 305–330). IGI Global Scientific Publishing.

[6] De Bois, M. et al. (2020). Enhancing the interpretability of deep models in healthcare using attention: A glucose forecasting application. *arXiv*, 2009.03732.

[7] Feature Contribution and Attention Mechanism-Based Explainability (2025). *Springer Nature Digital Medicine*.

[8] Li, X., Chen, H., & Wang, Y. (2025). Ensemble deep learning in healthcare: a systematic review. *Computers, Materials & Continua, 82*(3), 59945.

[9] Li, Y., Yao, X., & Padman, R. (2025). No Black Box Anymore: Temporal-Feature Cross Attention in clinical predictive modeling. *arXiv*, 2503.19285.

[10] Mahabub, S., Das, B. C., & Hossain, M. R. (2024). Advancing Healthcare Transformation: AI-Driven Precision Medicine and Scalable Innovations Through Data Analytics. *Edelweiss Applied Science and Technology, 8*(6), 8322–8332.

[11] Mahabub, S., Jahan, I., Islam, M. N., & Das, B. C. (2024). The Impact of Wearable Technology on Health Monitoring: A Data-Driven Analysis with Real-World Case Studies and Innovations. *Journal of Electrical Systems, 20*.

[12] Nasiruddin, M., Hider, M. A., Akter, R., Alam, S., Mohaimin, M. R., Khan, M. T., & Sayeed, A. A. (2024). Optimizing Skin Cancer Detection in the USA Healthcare System Using Deep Learning and CNNs. *The American Journal of Medical Sciences and Pharmaceutical Research, 6*(12), 92–112.

[13] Pant, L., Al Mukaddim, A., Rahman, M. K., Sayeed, A. A., Hossain, M. S., Khan, M. T., & Ahmed, A. (2024). Genomic Predictors of Drug Sensitivity in Cancer: Integrating Genomic Data for Personalized Medicine in the USA. *Computer Science & IT Research Journal, 5*(12), 2682–2702.

[14] Shah, P. et al. (2025). Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. *Scientific Reports, 15*, 17927.

[15] A stacking ensemble machine learning approach for the prediction of diabetes and gender-specific risks. (PMC11196524).

[16] Enhancing heart disease prediction with stacked ensemble and explainable AI. (2025). *Frontiers in Digital Health*.

[17] Zeeshan, M. A. F., Mohaimin, M. R., Hazari, N. A., & Nayeem, M. B. (2025). Enhancing Mental Health Interventions in the USA with Semi-Supervised Learning: An AI Approach to Emotion Prediction. *Journal of Computer Science and Technology Studies, 7*(1), 233–248.

[18] Zhang, Y., & Smith, J. (2025). Explainable attention in recurrent models for clinical time-series. *Journal of Medical AI, 2*(3), 145–157.

[19] Zhang, Y., Wang, F., & Chen, Y. (2022). Semi-supervised adversarial learning for mental health classification with imbalanced data. *IEEE Transactions on Affective Computing*.