An Explainable Machine Learning Approach to Personalized Healthcare Prediction

Rajeev Ranjan
Research Scholar, Department of Computer Science, V.K.S.University, Ara, Bihar

***Abstract*—A new paradigm in contemporary medicine, personalised healthcare seeks to tailor diagnosis, treatment, and preventative measures to each patient by analysing their unique set of medical history and other personal factors. While healthcare organisations have benefited greatly from Machine Learning (ML) integration in terms of improved predictive skills, many ML models' complexity makes them difficult to understand and use. This research presents a method for personalised healthcare prediction using Explainable Machine Learning (XML) that integrates both high predictive accuracy and model interpretability. Health outcomes and illness risks may be predicted using the framework by analysing patient demographics, clinical data, test findings, lifestyle variables, and medical history. For prediction, we use state-of-the-art ML algorithms. To understand how the model makes its decisions, we incorporate explainability techniques like SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and feature importance analysis. Improved trust, responsibility, and clinical decision support are outcomes of the suggested method, which helps healthcare providers comprehend the elements impacting forecasts. The explainable model maintains openness and user interpretability while achieving dependable prediction performance, according to experimental assessment. These results demonstrate the promise of explainable AI for enhancing medical decision-making, patient outcomes, and personalised healthcare methods.**

***Index Terms*—Health Care Forecasting, XML-Based Explainable Machine Learning, Predictive Analytics, XAI-Based Healthcare Prediction, AI Models with Full Transparency**

I. INTRODUCTION

Explainable machine learning has been created in order to bridge the gap between predictive healthcare models that are accurate and simply comprehensible. For the most part, the objective is to construct systems that are capable of reliably predicting outcomes while also providing explanations that are clinically accessible. Improving the understandability of models is being accomplished via the use of an expanding number of technologies. Methods such as attention procedures, SHAP (SHapley Additive explanations), LIME (Local Interpretable Model-Agnostic Explanations), and feature relevance analysis are all included in this collection of techniques. Utilizing these strategies may allow you to get an understanding of how each input variable influences the final prediction. Because of this, the forecasts become more trustworthy, responsible, and relevant to actual healthcare situations that occur in the real world. Additionally, explainable models significantly increase the safety of patients, the capacity of patients and healthcare practitioners to collaborate on decision-making, and compliance with regulations. All of these aspects are significantly improved. As a result of the ever-increasing amount and complexity of healthcare data, there is a rising need for artificial intelligence systems that are both successful and ethical in their behavior. The development of machine learning models that are capable of providing explanations for healthcare forecasts is an essential initial step in the process of developing trustworthy, patient-centered smart healthcare systems. This is a significant advance in technical capabilities.

II. MODERN HEALTHCARE SYSTEMS AND THE IMPORTANCE OF AI

When it comes to imaging in the medical field, artificial intelligence truly shines when it comes to medical imaging. As a result of the use of contemporary algorithms, it is now possible to correctly scan X-rays, CT scans, and MRIs for the presence of malignant tumors. Not only does this make

the work of radiologists simpler, but it also raises the percentage of early detection, which is an essential component of successful treatment outcomes. Clinical trials and the development of new drugs are two other fields that are using artificial intelligence. The former assists researchers in locating potentially useful medication concepts, while the latter improves trial designs in order to save expenses and the amount of time spent on them. On the other hand, despite these advantages, artificial intelligence is not yet capable of completely replacing human healthcare staff. Ethical issues, data privacy, algorithmic opaqueness, and the need for regulatory approval are only some of the challenges that must be overcome. A further disadvantage of computers is that they do not possess the human characteristics that are essential for healthcare, such as empathy, judgment, and interaction. Because of this, artificial intelligence is leading to changes in the healthcare industry by automating some operations and making them more precise. However, it is probable that AI will function more as a complement to human expertise than as a replacement for it. As time goes on, it is anticipated that artificial intelligence will work together with professionals in the healthcare industry, therefore enhancing their capabilities and eventually leading to a healthcare system that is both more effective and more focused on patients.

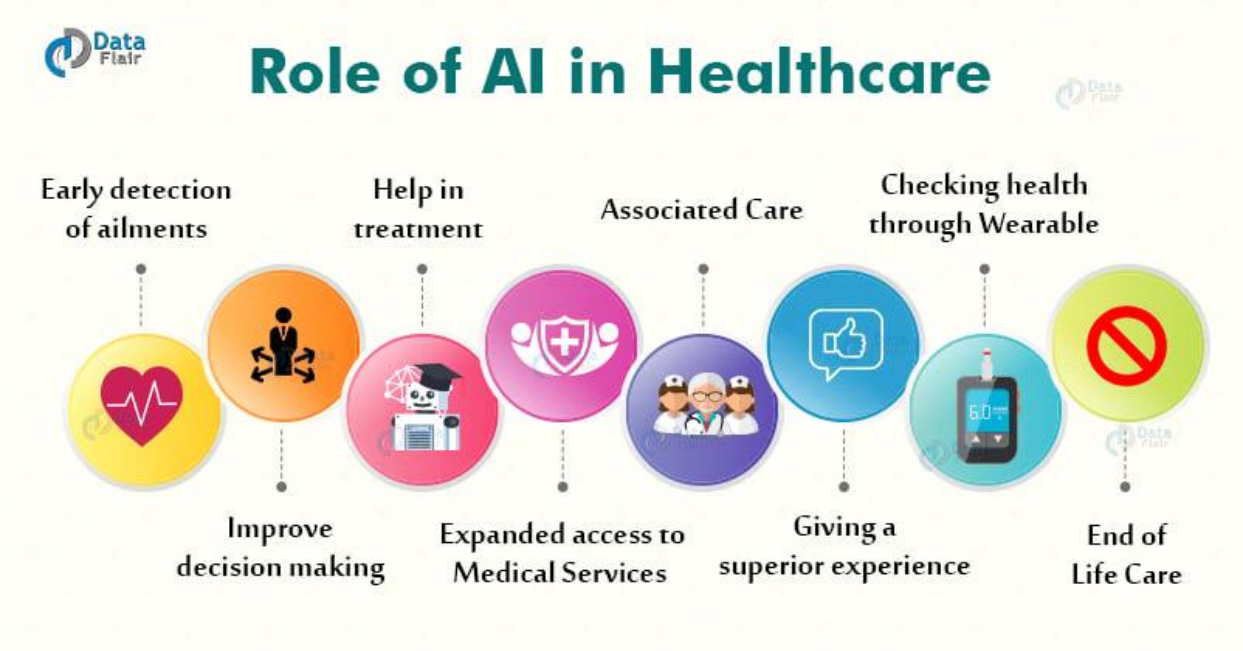

Figure.1. Applications and Importance of AI in Healthcare

III. REVIEWS

A. Medical Machine Learning

More sophisticated machine learning algorithms can now detect cardiovascular disease risk factors like hypertension, dyslipidemia, and bad lifestyle choices. In 2017, Krittanawong and colleagues demonstrated that deep learning models outperformed traditional risk score methods for predicting cardiac problems. Similarly, machine learning methods such as random forests and support vector machines (SVMs) have been successfully used to classify diabetic people according to their glucose levels, BMI, and family history. This led to the development of early intervention strategies (Suthaharan, 2016).

Machine learning has shown promising results in many important domains, one of which is cancer prediction. Models trained using machine learning to identify breast cancer have shown excellent sensitivity and specificity when applied to image data, such as scans. Litjens et al. (2017) found that convolutional neural networks (CNNs) greatly simplify the process of cancer detection in x-ray pictures. Customized cancer treatment is now within reach, thanks to machine learning-based genome research that has identified gene expression patterns linked to tumor

growth. The spread and occurrence of infectious diseases may be better predicted with the use of M.L. Predicting infection rates, hospital admissions, and fatality rates during the COVID-19 pandemic was a common task for machine learning algorithms. Legislators may benefit from machine learning-based prediction systems, as shown by Wang et al. (2020), when deciding how to allocate funds and keep problems under control.

In order to detect early warning signs of a patient's worsening condition, machine learning algorithms are increasingly being used. It is possible to identify sepsis, cardiac arrest, and organ failure in patients many hours before their symptoms worsen, according to models trained on datasets from intensive care units, such MIMIC (Medical Information Mart for Intensive Care). Machine learning models trained on data from critical care units may help doctors spot signs of sepsis earlier, which might lead to fewer fatalities (Johnson et al., 2016).

Healthcare as a whole has benefited greatly from clinical decision support systems (CDSS) driven by machine learning as they help doctors make better judgment calls. These systems provide evidence-based suggestions for the diagnosis, treatment, and management of diseases based on patient data. When properly integrated into healthcare delivery, clinical decision support systems (CDSS) improve clinical outcomes, say Kawamoto et al. (2005).

Electronic health records (EHRs), laboratory findings, imaging data, and patient histories are just some of the data kinds that the ML-based CDSS can handle. As a result of combining all of these different types of data, machine learning models help doctors better understand their patients' complicated diseases. For example, cancer researchers utilize rule-based and mixed machine learning algorithms to tailor treatment plans to each patient by taking into account their unique disease and symptoms. The use of ML-powered CDSS might be useful for radiologists as it helps them interpret medical pictures and spot abnormalities. Esteva et al. (2017) demonstrated that deep learning models can accurately categorize skin cancer to the same degree as a doctor, suggesting that these models might be useful for diagnostic decision-making. Machine learning algorithms work in a similar way as pathologists, assisting them in detecting problems in tissue samples. This speeds up the diagnostic process and helps avoid mistakes.

B. Predictive Healthcare Models

Heuristics are used by traditional Early Warning Systems, such as the Modified Early Warning Score (MEWS) and the National Early Warning Score (NEWS), in order to ascertain whether or not an individual's health is worsening throughout the course of the assessment. This is accomplished by these systems by adding up the vital indications of a person, which include their heart rate, blood pressure, and breathing rate, among other things. The most notable of these traditional scoring systems' limitations is that they are inflexible and are unable to explain how parts interact in ways that are intricate and non-linear. Despite the fact that they are straightforward, these scoring systems have a number of disadvantages. According to Smith et al. (2019), they are also susceptible to errors that are produced by humans while they are doing calculations and writing down the findings.

EWS has been dramatically transformed by machine learning, which has made it possible to automatically and continuously classify risks. Recent studies have shown that machine learning-based emergency warning systems are superior in their ability to anticipate potentially catastrophic events, such as cardiac arrest, unexpected transfers to the intensive care unit, and activations of rapid response teams, many hours before the typical clinical triggers occur. Using data from electronic health records (EHR) in deep learning models, predicting clinical deterioration across many time periods with much greater Area Under the Receiver Operating Characteristic Curve (AUROC) scores than typical NEWS2 methods was shown by Rajkomar et al. (2018). Here is just one instance. Furthermore, studies on an EWS based on continuous machine learning by Kipnis et al. (2016) shown that these models could accurately predict unfavorable events up to 48 hours in advance. This allowed medical personnel more time to respond and provide aid.

Methods used in this field have progressed from more basic supervised learning algorithms (like Random Forests and Logistic Regression) to more complex temporal models. Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) are two popular options because they adapt to the ever-changing nature of patient deterioration. With these efficient methods, time-series correlations in vital signs may be found. They may notice trends, such a

slow rise in breathing rate followed by a little drop in blood pressure, for example (Churpek et al., 2016). However, despite its superiority in prediction quality, EWS based on machine learning has a hard time gaining adoption. The term "alert fatigue" appears often in academic articles. Research from 2020 by Sendelbach et al. suggests that this happens when overly sensitive models produce a high volume of false positives, leading nurses to become less vigilant. This is why new studies recommend using both calibration measures (like the Brier score) and classification metrics (like AUROC) simultaneously when building models. Making sure the odds of anything happening are close to the rates of real clinical deterioration is the reason for this.

A potentially fatal organ failure known as sepsis develops when the host's immune system reacts abnormally to an infection. All across the world, it is still the leading killer of hospitalized people. Seymour et al. (2017) found that there is a 7.6% increase in mortality for every hour that goes by without giving broad-spectrum antibiotics to patients with sepsis, highlighting the critical need of acting quickly when treating this condition. One way to diagnose sepsis has been using the SIRS criteria or the Sequential Organ Failure Assessment (SOFA) score, both of which have been around for a long time. Contrarily, these measures are often reserved for situations when a clinical suspicion of infection is available. This means that the vital chance to undergo minor physiological changes during the early stages is gone.

Critical care analytics has taken a giant leap forward with the use of machine learning models to sepsis detection. There has been a shift in thinking about how to utilize EHR data from both general wards and critical care units, with examples coming from the MIMIC-III and MIMIC-IV databases. This action is being taken in order to foresee the onset of sepsis hours prior to its clinical manifestation. Nemati et al. (2018) made a groundbreaking contribution to the field of artificial intelligence-based sepsis prediction systems. Their work includes the creation of the TREWS algorithm, which makes use of deep learning models and a set of gradient-boosted trees. In comparison to previous baselines, their model demonstrated a much better area under the receiver operating characteristic curve (AUROC) of 0.85 and, more importantly, a median early lead time of 4.5 hours before the start of sepsis.

The condition known as sepsis is very difficult, which is why we need sophisticated machine learning techniques. Researchers have repeatedly discovered that sepsis does not follow a single, linear biochemical path. This is something that has been confirmed to be the case. This demonstrates that linear models are not always successful. Based on the findings of Alistarh et al. (2021), it has been shown that deep learning models, particularly Long Short-Term Memory (LSTM) and Temporal Convolutional Networks (TCNs), are capable of mapping the complex and diverse pathways that sepsis takes. In addition, research that has recently been conducted has used both random data and ordered data from laboratories and vital signs. When natural language processing models examine clinical nursing notes and physician records, they are able to identify early subjective complaints (such as "patient seems confused" or "complains of chills") that occur many hours prior to the occurrence of objective test abnormalities (Goh et al., 2021).

There is still a great deal of controversy about whether or not sepsis prediction models are accurate in the actual world, despite the fact that major advancements have been made. A single center is responsible for the training and testing of a large number of high-quality models. According to Mandel et al. (2021), this may result in algorithmic bias and poor generalizability when it is used in a variety of clinical settings with a large number of distinct patient populations and electronic health record systems. In addition, depending on laboratory data results in a timing bias since the data are often collected at the time when a physician suspects the presence of an illness. In order to address this issue, recent research emphasizes the need of prospective studies that are conducted at several sites and the development of models that only make use of data that is readily accessible at the bedside (such as demographics and vital signs) in order to guarantee that screening is both fair and efficient.

When attempting to predict who will pass away in the intensive care unit (ICU), machine learning takes advantage of the vast volumes of heterogeneous data that are generated in these settings. These data include continuous signals (such electrocardiograms), discrete lab results, records of medicine distribution, and clinical activities. ICU fatalities may be identified more accurately by machine learning methods than by

traditional severity ratings, according to the findings of a large number of systematic reviews and meta-analyses. With regard to categorization, for example, Pirracchio et al. (2015) demonstrated that Random Forest models that used EHR data performed much better than SAPS II and APACHE IV approaches. After then, research that used deep learning on temporal data brought the AUROC value up to a level that was more than 0.90. This demonstrated that artificial intelligence is capable of discovering intricate, non-linear relationships between organ systems, which are difficult to comprehend for medical professionals or conventional statistical models (Hug et al., 2021).

The transition from static, admission-only predictions to dynamic, updated risk scores represents a significant advancement in the process of forecasting mortality in intensive care units. Because of the rapid pace at which a patient's condition changes in the intensive care unit (ICU), a fixed 24-hour score is soon rendered obsolete. "Rolling" forecasts of death are currently being made using LSTM networks and attention-based transformer models (Harutyunyan et al., 2019). These predictions are updated every hour depending on fresh information that is gathered. Because of this, medical professionals are able to determine immediately whether or not the therapies they are administering are effective. If the patient's anticipated mortality risk decreases following a fluid bolus or vasopressor modification, the model verifies that the patient is on the correct path.

On the other side, the incorporation of machine learning death models into medical procedures would result in significant ethical and practical challenges. When it comes to circumstances that include end-of-life care, the "black box" nature of very complex deep learning models presents a significant challenge. According to Amann et al.'s research from 2020, clinicians do not want to switch methods of treatment, such as transitioning to comfort care, because of a risk number that they are unable to comprehend or evaluate. Explainable artificial intelligence (XAI) is the subject of a significant amount of recent study when it comes to forecasting mortality in the intensive care unit (ICU). The SHapley Additive Explanations (SHAP) technique is being used in these models in order to determine which elements (such as increasing lactate levels and decreased urine output) are responsible for a high mortality estimate at any particular hour (Lundberg et al., 2020). According to the majority of the study, in order for intensive care unit death models to be used in real life, they need to evolve from only being predictive tools to being able to assist medical professionals in making choices.

C. Class Imbalance Problem in Healthcare

The assumption that the classes in the dataset are somewhat uniformly distributed is the foundational premise upon which the bulk of machine learning algorithms are created. The term "class mismatch" refers to the situation in which one group constitutes a disproportionately high percentage of the population in comparison to another group. Rather than being the result of inadequate data collection, Johnson and Khoshgoftaar (2019) assert that this bias is really a characteristic of epidemiology in the healthcare business. The majority of the time, there are fewer people who are sick than there are healthy ones, which is why medical records are often lopsided.

This mismatch can be put into a range of categories based on how bad it is. Mild imbalance could mean that there are 1 in 3 or 1 in 5 minority cases compared to majority cases. This happens a lot in certain age groups with chronic diseases like Type II Diabetes. Extreme mismatch, with ratios of 1:100, 1:1000, or higher, is the norm for early disease discovery, rare genetic illnesses, and predicting bad things that will happen (Fernández et al., 2018). In a general medical ward, for example, most of the observations made on each patient hourly will be in the "non-sepsis" class, while the real change from "non-sepsis" to "sepsis" will only make up a very small part of the dataset.

The mismatch between classes has a significant impact on machine learning models since it causes them to learn more in favor of the class that most people belong to. Standard machine learning techniques include Support Vector Machines (SVMs), Random Forests, and Deep Neural Networks are designed to enhance global goal functions, most often the accuracy of classification (Weiss, 2004). These algorithms are used to increase the accuracy of classification. When these algorithms are presented with an imbalanced dataset, they determine that the process of ignoring the minority class and always guessing the majority class is the most efficient and accurate approach to get everything correctly.

A lot of people have written about this occurrence, which is known as the "accuracy paradox." One

example is a dataset where 99% of patients do not have a rare heart condition. A simple model that guesses "healthy" for all of those patients will still be 99% right. But this kind of model isn't useful in real life because it never correctly predicts the state it was made to find (Japkowicz & Stephen, 2002).

In addition to the accuracy paradox, class mismatch messes up the decision limits that classifiers learn in a big way. To separate classes, algorithms use the physical and mathematical features of the data. There are a lot of false positives when there aren't many cases of the minority class. This is because the decision limit is often twisted or pushed too close to the minority class. Also, datasets that aren't fair often have features and noise that overlap. Since there aren't many examples of the minority class, a small amount of noise or unusual data can change how that class is learned in a way that isn't fair (Weiss, 2004). In complicated, high-dimensional healthcare data, this makes models that don't transfer, showing a lot of variation and shaky performance when they are put to use on new groups of patients. It's hard for even the most advanced deep learning architectures to deal with extreme imbalance. This is because the majority class often dominates the gradient during backpropagation, which makes the network's weights converge to a local minimum that ignores the minority signal (Johnson & Khoshgoftaar, 2019).

When it comes to professional practice, the costs of wrongly classifying someone are very uneven. A fake positive, which means that a test finds a disease when there isn't one, usually causes extra stress and often expensive diagnostic treatments (for example, a biopsy after a false-positive mammogram). Even though this mistake is bad, it can usually be fixed. On the other hand, a false negative not finding a patient who actually has the disease delays treatment that is needed, lets the disease get worse, and can lead to avoidable illness or death (Rajpurkar et al., 2017). Class imbalance is a direct threat to the "first, do no harm" ethical requirement of medicine because it forces models to produce fake positives in order to keep total accuracy.

In addition, fixing the mismatch between classes is very important for the main goal of this thesis, which is to create Explainable Machine Learning (XAI). It is natural for the model's decision logic to get more complicated when researchers use methods to fix class mismatch, like synthetic minority oversampling (SMOTE) or cost-sensitive learning. To make sure the right classification is made, algorithms trained on resampled data or unequal punishment structures learn to find very specific, localized patterns in the minority class (Lundberg et al., 2020).

If a model has been strongly tuned to fix a huge class mismatch, doctors can't trust its results until they know why it was changed in the first place. In this situation, being able to explain something becomes a clinical necessity. For instance, if a XAI model suggests that a patient will get a rare type of cancer and the explainability module shows that the prediction was based on a real but slight interaction between certain biomarkers and not an error caused by oversampling, the doctor can be sure that the forecast is correct and take action. If models that try to get around class imbalance can't be explained, they are often seen as "black boxes" that send out random alerts, which can lead to alert fatigue and, eventually, rejection by medical staff (Amann et al., 2020).

D. Techniques for Handling Class Imbalance

The simplest way to deal with class imbalance is by Random Under Sampling, or RUS for short. To attain the target minority class balance ratio, it randomly removes instances from the majority class until it achieves the target level. The most direct route to resolving the problem of socioeconomic inequality is RUS. Reduced training time and memory overhead are two benefits of RUS's method of drastically paring down dataset sizes. Weiss (2004) suggests that this may be very helpful for training complex deep learning architectures on massive EHR datasets. The computational efficiency of RUS is the most essential advantage when it comes to benefits.

The medical literature, however, is quite clear that RUS should not be used in fields that fail to adequately account for relevant factors. One of the most important and significant problems with RUS is the possibility of data loss. People who are classified as "healthy" or "non-event" constitute the bulk of the population in healthcare statistics. Physiological sub-phenotypes and changes that are noticeable at baseline level fall under this group. Underfitting might occur if these cases are rejected arbitrarily, since crucial boundary information could be lost. The real decision border that distinguishes healthy states from sick ones may slip the mind of a model trained on RUS data, resulting in subpar generalization. If the model had been trained

with RUS data, this would have been the result. When used to high-dimensional clinical data, RUS often affects the performance of complicated models, according to study by Yoon and Kim (2020). Reason being, abnormalities are often uncovered by comparing the rejected instances to the normal physiological spectrum, which typically contains tiny patterns. For this reason, this remains true.

Despite its ease of use, ROS has taken a lot of heat in the literature on predictive modeling. One of the main culprits behind overfitting phenomena is reactive oxygen species (ROS). Machine learning algorithms, especially neural networks and decision-tree models, tend to remember these individual repeated samples instead of generalizing the underlying clinical theory, as stated by Batista et al. (2004). Given that these algorithms are built to learn from data, this poses a challenge. The reason for this is because the training set includes case studies that mirror the experiences of marginalized communities. Although this patient's biomarker profile was successfully identified several times in the training set, the model will not be able to handle a slightly altered phenotypic presentation of the same illness. This is a crucial consideration when making a medical diagnosis. Furthermore, ROS incorrectly emphasizes the significance of the replicated samples, leading to a skewed statistical distribution that does not provide anything further of value. Because of this, it seems like the model is sturdier than it really is, which isn't entirely true.

To shorten its term, SMOTE stands for Synthetic Minority Over-sampling Technique. To get around ROS's overfitting restrictions, Chawla et al. (2002) developed the Synthetic Minority Over-sampling Technique, or SMOTE. One of the most well-known and widely-used resampling algorithms in healthcare analytics, this approach has stood the test of time. For the purpose of creating fresh instances of synthetic minority groups, SMOTE employs the interpolation approach. Compared to the technique of making exact copies of existing instances, this is different. A process called SMOTE finds the minority sample's nearest neighbors in the feature space, then randomly chooses one or more of them. The last step is to generate new data points along the line segments that connect the original sample to its neighbors. This procedure is carried out again and again until the required quantity of data points is produced.

By extending the minority decision area, SMOTE introduces a revolutionary shift in the learning paradigm. The learning paradigm is radically changed as a result of this change in emphasis from memorization to generalization. According to Fernández et al. (2018), SMOTE might potentially enhance sensitivity (recall) in clinical applications like early-stage cancer diagnosis or the prediction of uncommon cardiovascular events without substantially lowering specificity.

Regardless, there are substantial limitations associated with using SMOTE in healthcare data systems. It is definitely within the realm of possibility for SMOTE to generate synthetic patients who defy logic or science. This happens because it functions in the feature space, completely oblivious to the clinical setting in which the variables are located. Instances of behaviors that breach domain limitations include creating a synthetic "age" of 150 years and interpolating between a "male" and a "female" patient (Han et al., 2005). You may say that both of these things are in violation of domain regulations. Class overlap is already a major problem, yet the basic SMOTE method doesn't even consider it. When it comes to medical care, those who are already ill and those who are well typically inhabit very identical environments. There may be a lot of overlap, for example, between an elderly patient with a few comorbidities and a young patient in the beginning stages of the disease. In these overlapped areas, SMOTE may potentially create synthetic samples. It would become more challenging to explain future occurrences due to the introduction of noise into the majority class space and the muddying of the choice boundary.

Even though SMOTE and ADASYN have made tremendous progress in recent years, hybrid techniques are increasingly being recommended as the most effective way of getting adequate model calibration. This is because hybrid methods combine hybrid approaches with traditional methods. The use of hybrid methods, which combine oversampling with targeted under sampling or data cleaning, may be able to reduce the boundary blurring and noise production that are inherent in synthetic approaches. This is because hybrid methods combine the two techniques. According to Batista et al. (2004), two notable hybrid techniques are SMOTE-ENN and SMOTE-SMOTE Tomek, which combine SMOTE with Edited Nearest Neighbors. The combination of both of these methods

is called SMOTE. You can get a lot of use out of both of these hybrid approaches. Synthetic sample generation is the first stage of these paradigms. After that, instances that are suspiciously near the other class are either identified and removed using ENN techniques or Tomek connections. Iterations of this process are performed until the targeted results are achieved. There is now a great deal more distinction between healthy and sick states as a consequence of this thorough cleansing of the decision boundary. In clinical prediction tasks, recent comprehensive evaluations in medical informatics have shown that hybrid systems often outperform solo SMOTE or ADASYN. This holds truest when ensemble classifiers such as XGBoost or Random Forests are used with hybrid approaches (Guo et al., 2020). The two together are very efficient.

IV. MACHINE LEARNING ALGORITHMS USED IN HEALTHCARE

The operation of Decision Trees (DT) is accomplished by means of a recursive and hierarchical division of the feature space. This partitioning results in the creation of a series of boolean if-then rules, which finally result in a clinical prediction. When it comes to complicated models, DTs are often regarded as the "gold standard" because of their capacity to describe complex phenomena. Their graphical, tree-like structure closely mimics human clinical reasoning and established medical guidelines (for example, "If patient age is greater than 65 and troponin levels are greater than X, then diagnose myocardial infarction"), which is one of the reasons why they are highly favored in clinical decision support systems (Freeman et al., 2020).

Standalone DTs, on the other hand, are notoriously unstable and often subject to substantial fluctuation. Even very slight changes to a clinical training dataset, such as the inclusion of a few patient records that are considered to be outliers, may result in an entirely different tree structure, as shown by the research that has been conducted. In addition, deep learning models have a significant tendency to overfit, which is especially problematic when they are trained on high-dimensional healthcare data that does not have strong depth limits. The data will be fragmented into extremely specialized leaf nodes by an overfitted DT, which will achieve perfect training accuracy but will fail catastrophically to generalize to patient groups that have not been observed before, hence making the rules that seem to be intuitive clinically incorrect (Lipton, 2018).

The stacked ensembles represent the pinnacle of the "black-box" dilemma, despite the fact that they typically attain the highest possible AUROC scores in clinical predictive modeling. Due to the fact that the final forecast is a combination of numerous intricate models, the inherent interpretability of the prediction makes it entirely impossible. the need for artificial intelligence that is post-hoc and explainable. Frameworks for model-agnostic explainability, such as SHAP or LIME, are required to be used in order to guarantee that a stacked ensemble may be utilized for therapeutic purposes. According to the research that was carried out by Lundberg and colleagues in the year 2020, these solutions break down the complicated ensemble output into the localized and additive contributions of the distinct clinical factors that are associated with each individual patient.

V. FEATURE ENGINEERING IN HEALTHCARE DATA

Because of the nature of the information that pertains to healthcare, it is not possible to completely eradicate missing data. When it comes to therapeutic contexts, missing numbers are often fairly useful, in contrast to some other areas where the lack of these figures indicates that an error has occurred. In most cases, the absence of a troponin level or any other laboratory result indicates that the practitioner did not believe that a cardiac event was taking place. Based on this information, it may be deduced that the lack of the test is directly related to the current state of the patient's health. Previous research has shown that there are three unique ways in which data might be missing: missing completely (MCAR), missing randomly (MAR), and missing not at random (MNAR). Each of these three types of missing data is presented below. EHR data is where MAR and MNAR trends are most often seen, according to the findings of Little and Rubin (2019).

The use of listwise deletion, which involves the removal of patients whose records are incomplete, or any other naïve approaches to address the issue of missing data is strongly discouraged in the field of medical machine learning. The reason for this is

because these approaches have the potential to result in very tiny samples and a significant amount of selection bias. According to Che et al. (2018), this is especially true in situations when the most critically ill patients are more likely to have dynamic variables overlooked due to interruptions in the intensive care unit. In contrast, straightforward imputation techniques, such as mean or median imputation, are often used because of the ease with which they may be applied. The research literature, on the other hand, cautions that mean imputation might falsely lower the variance of the clinical variable and change the distribution that is underlying the data. It is possible that this may lead downstream explainability modules to become exceedingly confused, which in turn may cause scientists to make significant errors when interpreting their data. This is because true physiological extremes will be disguised.

In the field of healthcare statistics, outliers provide a significant methodological challenge. When people think about machine learning, they often conceive of outliers as random noise or faulty measurements that bias model training. However, outliers are not random noise. In the field of statistical outlier treatment, some of the methods that are often used include Z-score truncation and IQR capping, amongst others. A physiological outlier, on the other hand, is not always an error in the clinical data; rather, it is the specific physiological signal that suggests a potentially deadly clinical event, such as a cardiac arrhythmia or severe renal failure (Jiang et al., 2021). One example of a physiological outlier is a potassium level of 7.0 mEq/L or a heart rate of 180 beats per minute. Both of these values are considered abnormal.

Because of this, the normal outlier removal procedures that are used in predictive healthcare processes should not be implemented without proper consideration; they might potentially have devastating consequences. The predictive signal that is associated to critical care treatments and severe deteriorations is effectively disrupted as a result of these high values being controlled or abolished. It has been shown by research that clinical domain expertise, as opposed to relying only on statistical criteria, might have a substantial influence on outlier management in the healthcare industry. According to the findings of the research carried out by Le et al. in 2022, it is essential to keep in the dataset any outliers that are found to be indicative of genuine physiological extremes, as opposed to artifacts that are the result of sensors or human mistake in typing.

The process of changing continuous clinical information into discrete, categorical bins is referred to as discretization, which is also known as batching. Within the realm of healthcare, this method is profoundly ingrained in clinical practice. Physicians seldom conceive of a patient's systolic blood pressure as a continuous float (for example, 142.5 mmHg), but rather in categorical terms (for example, "normal," "elevated," or "hypertensive crisis").

The research that has been done has identified a few different discretization approaches. In unsupervised approaches, the data range is divided mathematically without taking into consideration the outcome variable. Examples of such methods are equal-width binning and equal-frequency binning. According to García et al. (2013), they are often criticized in the field of medical machines since they have the ability to put clinically diverse states into the same bin. Methods of supervised discretization, such as Decision Tree-based binning (for example, utilizing a shallow tree to discover optimum split points that maximize information gain relative to the target illness), are preferable from a mathematical standpoint. Domain-driven discretization, on the other hand, is generally considered to be the most reliable method in explainable healthcare. As an example, the official ADA thresholds for HbA1c levels are used to define diabetes status, and the SIRS criterion thresholds for heart rate and temperature are utilized (Kuo et al., 2021). This entails the binning of continuous variables in accordance with published medical recommendations.

VI. ARTIFICIAL INTELLIGENCE THAT CAN BE EXPLAINED (XAI)

It is very challenging to use general XAI principles in healthcare settings due to the inherent asymmetry that exists within clinical decision-making. This undertaking is very difficult to do. When applied to the field of medicine, the costs of mistakes are quantified in terms of the morbidity and death of patients. As a consequence of this, the significance of XAI in the field of healthcare is founded on the idea of "appropriate trust." In the event that a model suggests beginning a high-risk intervention, such as the administration of a powerful vasopressor or the

execution of mechanical breathing, a physician cannot proceed on the basis of blind trust. According to Amann et al.'s research from 2020, they are required to examine the advice together with their clinical competence, the patient's medical history, and established medical criteria.

Additionally, the ideals of autonomy and informed consent are governed by the principles of medical ethics. Patients have an ethical right to an understanding of the elements that influence their treatment plans, particularly as machine learning models continue to play an increasingly important role in diagnosis and prognosis. (Cabitza et al., 2017) Through the process of converting abstract computational probabilities into narratives that are clinically relevant, XAI makes it easier for several people to make decisions together. The output of the artificial intelligence is aligned with clinical causation when, for instance, it is explained that a patient's increased risk for acute kidney damage is caused by the recent administration of a particular nephrotoxic medicine and a decreasing baseline creatinine level. This empowers both the physician and the patient.

Model-agnostic, post-hoc explanation strategies have become more prevalent in the research literature as a means of operationalizing interpretability for complicated mathematical models. Both SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are the two algorithms that dominate this domain.

The LIME algorithm, which was first presented by Ribeiro et al. (2016), provides an explanation for individual predictions by roughly approximating the black-box model locally with an interpretable surrogate model, which is often a linear regression or decision tree. As part of its operation, LIME involves perturbing the input data of a particular patient (for example, by making tiny adjustments to their laboratory results), monitoring how the prediction of the black-box model changes, and then fitting a straightforward model to the perturbed samples. Despite the fact that LIME is very user-friendly and frequently used, the research that has been done on it reveals that its application to healthcare has considerable limits. Due to the fact that LIME is dependent on random perturbations, it has the potential to produce patient profiles that are not realistic (for example, a patient who has both hyperkalemia and hypokalemia at the same time). In addition, LIME explanations have the potential to be unstable; if the algorithm is performed numerous times on the same patient, it may provide varied feature attributions, which might undermine clinician confidence (Zhou et al., 2021).

When it comes to researching artificial intelligence in healthcare settings, the SHAP framework has become the method of choice. Cooperative game theory and, more particularly, the Shapley values, which were introduced by Lloyd Shapley in 1953, are the foundations upon which it is built. Within the context of machine learning, Lundberg and Lee (2017) made adjustments to Shapley values in order to make them more applicable. They identified the precise marginal contribution of each attribute for each prediction that they made. SHAP is a mathematical property that guarantees a variety of desirable mathematical features. The local accuracy, the missingness, and the consistency are some examples of these. In the event that a model develops in such a manner that a feature's contribution increases, the SHAP value of that feature should not decrease. Furthermore, it is essential that the total of the SHAP values corresponds to the output of the model.

VII. CONCLUSION

A major step forward for contemporary healthcare systems has been the creation of an explicable machine learning method for individualised healthcare prediction. This model not only accurately predicts illness risk, treatment results, and patient health status by combining predictive analytics with explainability methodologies, but it also clearly reveals the elements that influence these predictions. Healthcare providers and patients can have more faith in AI-driven decisions when there is this level of openness.

This research proves that explainable ML models may enhance evidence-based clinical decision-making by efficiently analysing complicated healthcare data without losing interpretability. By determining which risk variables are most important and how they affect predicted results, healthcare professionals may create individualised plans for prevention and focused treatments. In addition, by offering patients clearer reasons for predictions, explainable AI has the potential to enhance diagnosis accuracy, optimise resource utilisation, and promote improved patient participation in healthcare. In spite of problems with

privacy, model generalizability, and data quality, the results show that explainable machine learning might be a good tool to improve individualised healthcare. Finally, explainable machine learning provides a solid foundation for building predictive healthcare systems that are open, precise, and focused on the needs of individual patients. Integrating explainability approaches into real-world clinical settings, enhancing model robustness, and including varied healthcare datasets should be the focus of future research to maximise their influence on healthcare outcomes and patient well-being.

REFERENCES

[1] Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[2] Fernández, A., S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," J. Artif. Intell. Res., vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[3] Guo, H., Y. Li, J. Shang, M. Gu, Y. Huang, and B. Gong, "Learning from class-imbalanced data: Review of methods and applications," Expert Syst. Appl., vol. 168, Art. no. 114327, 2021, doi: 10.1016/j.eswa.2020.114327.

[4] Han, H., W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Advances in Intelligent Computing, Berlin, Germany: Springer, 2005, pp. 878–887, doi: 10.1007/11538059_91.

[5] He, H., Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in Proc. IEEE Int. Joint Conf. Neural Netw., 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[6] Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee, "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell., vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.

[7] Weiss, G. M., "Mining with rarity: A unifying framework," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7–19, 2004, doi: 10.1145/1007730.1007734.

[8] Yoon, S., and J. Kim, "Data-level approaches for imbalanced biomedical data classification," Appl. Sci., vol. 10, no. 15, Art. no. 5033, 2020, doi: 10.3390/app10155033.

[9] Zhang, X., and Y. Li, "Adaptive synthetic oversampling combined with ensemble learning for imbalanced medical data classification," J. Healthcare Eng., vol. 2021, Art. no. 6612734, 2021, doi: 10.1155/2021/6612734.

[10] Ahmad, M. A., A. Teredesai, and D. Turaga, "Explaining black box models with transparent, multi-stage surrogate models," in Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2018, pp. 1329–1338, doi: 10.1145/3219819.3219933.

[11] Breiman, L., "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[12] Chen, T., and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[13] Díaz-Uriarte, R., and S. A. De Andres, "Gene selection and classification of microarray data using random forest," BMC Bioinformatics, vol. 7, no. 1, Art. no. 3, 2006, doi: 10.1186/1471-2105-7-3.

[14] Freeman, R., A. Vashistha, and R. Anderson, "An audit of algorithms for predicting emergency admissions: Implications for clinical decision support," J. Am. Med. Inform. Assoc., vol. 27, no. 10, pp. 1574–1581, 2020, doi: 10.1093/jamia/ocaa126.

[15] Freund, Y., and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, 1997, doi: 10.1006/jcss.1997.1504.

[16] Kourou, K., T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[17] Kwon, D., H. Kim, Y. J. Kim, S. J. Choi, J. E.

Lee, and S. Park, “Clinical application of logistic regression and machine learning models for risk prediction in healthcare,” Healthcare Inform. Res., vol. 27, no. 4, pp. 310–317, 2021, doi: 10.4258/hir.2021.27.4.310.

[18] Lipton, Z. C., “The mythos of model interpretability,” Queue, vol. 16, no. 3, pp. 31–57, 2018, doi: 10.1145/3236386.3241340.

[19] Luts, J., A. Molinaro, J. De Brabanter, K. De Wolf, J. A. K. Suykens, and S. Van Huffel, “A comparison of SVM and LS-SVM for classification in medical applications,” IFAC Proc. Vol., vol. 43, no. 11, pp. 44–49, 2010, doi: 10.3182/20100826-3-TR-4012.00009.

[20] Nemati, S., A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, “An artificial intelligence system for predicting the onset of sepsis in the ICU,” Nat. Med., vol. 24, no. 11, pp. 1706–1711, 2018, doi: 10.1038/s41591-018-0213-y.

[21] Rajkomar, A., E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al., “Scalable and accurate deep learning with electronic health records,” NPJ Digit. Med., vol. 1, no. 1, Art. no. 18, 2018, doi: 10.1038/s41746-018-0029-1.

[22] Che, Z., S. Purushotham, B. Khemani, Y. Liu, and D. Sontag, “Recurrent neural networks for multivariate time series with missing values,” Sci. Rep., vol. 8, no. 1, Art. no. 6085, 2018, doi: 10.1038/s41598-018-24271-9.

[23] García, S., J. Luengo, J. A. Sáez, V. López, and F. Herrera, “A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 4, pp. 734–750, 2013, doi: 10.1109/TKDE.2012.35.