IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 21 February 2023, accepted 24 March 2023, date of publication 27 March 2023, date of current version 31 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3262467

RESEARCH ARTICLE

Development of an Artificial Intelligence-Supported Hybrid Data Management Platform for Monitoring Depression and Anxiety Symptoms in the Perinatal Period: Pilot-Scale Study

NUR BANU OĞUR[1], CELAL ÇEKEN[1], YAVUZ SELİM OĞUR[2], HİLAL USLU YUVACI[3], AHMET BÜLENT YAZICI[2], AND ESRA YAZICI[2]

[1]Department of Computer Engineering, Faculty of Computer and Information Sciences, Sakarya University, 54050 Sakarya, Turkey
[2]Department of Psychiatry, Faculty of Medicine, Sakarya University, 54050 Sakarya, Turkey
[3]Department of Obstetrics and Gynecology, Faculty of Medicine, Sakarya University, 54050 Sakarya, Turkey

Corresponding author: Nur Banu Oğur (nbogur@sakarya.edu.tr)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Sakarya University Research Hospital under Approval No. E-71522473-050.01.04-112781-38.

ABSTRACT One of the forces driving science and industry is machine learning, but the proliferation of Big Data necessitates paradigm shifts from conventional approaches in applying machine learning techniques to this massive amount of data with varying velocity. Computers are now capable of accurately diagnosing a variety of medical conditions thanks to the availability of immense healthcare datasets and advancements in machine learning techniques. The study's primary aim is to identify the most compelling questions on anxiety and depression in pregnant women by extracting features through performance-optimized algorithms. In this way, it is aimed to reach the result in a shorter time with fewer questions. The next goal of this work is to create an instant remote health status prediction system for depression and anxiety in pregnant women based on the Apache Spark Big Data processing engine, which concentrates on using machine learning models on streaming Big Data. In this scalable system, the application receives data from pregnant women to forecast the patient's health condition. It then applies the Naïve Bayes machine learning algorithm that produces the best results for this dataset with accuracy and precision 90.8% and 81.71% respectively. With the assistance of this big data platform, the time-consuming anxiety and depression detection procedure in a pregnant woman can be replaced with a computer-based technique that works in an instant with a respectable amount of accuracy.

INDEX TERMS Data analytics, streaming data processing, machine learning, health informatics, perinatal period, anxiety, depression.

I. INTRODUCTION

The current era, particularly the previous two decades, might be referred to as the "age of big data," in which digital data is becoming increasingly important in various disciplines, including research, technology, society, and healthcare [1]. These datasets provide significant obstacles to the computing resources and analytics framework by making the whole study challenging for quickly extracting valuable information. Therefore, creating an effective big data analytics framework is a crucial research issue to overcome these sorts of problems. There are intensively used a number of research

The associate editor coordinating the review of this manuscript and approving it for publication was Mansoor Ahmed.

works in creating a big data analytics framework, such as Apache Spark [2], Apache Hadoop [3], Apache Storm [4], and Apache Kafka [5], to solve healthcare problems. Mental health diseases are health problems that deeply affect the lives of individuals and must be treated with care. If the psychological disorders that appear in the perinatal period, which includes the pregnancy process, are not detected in time, they will have negative effects on both the mother and the baby. If this is the case, it is of great importance for the mental health of society.

Machine learning has the potential to inform illness models, the discovery and development of innovative medicines that can affect disease, as well as prevention methods in psychiatry [6].

Machine learning algorithms used to make sense of data on the big data analytics platform reveal results for the solution. In the healthcare field, it promises to inform the discovery and development of new disease-healing treatments. Techniques used in the field of recognizing brain and mental disorders have created incomplete or erroneous representations, and the results have not been sufficient. It is important for public health that psychological disorders benefit from the developments to be obtained by creating big data analytics. It can be observed that psychological disorders, which are health problems that deeply affect the lives of individuals and need to be treated carefully, are more likely to occur in some periods of life. The perinatal period, which includes the pregnancy process, is one of these processes. To mitigate the damage of this era, researchers and psychiatrists have an unprecedented chance to make use of intricate patterns in the brain, behavior, and disease utilizing machine learning techniques [7]. By using these analysis techniques, it is aimed to detect important psychiatric disorders early, identify related factors and develop preventive measures.

The rate of psychological disorders appearing in the perinatal period is quite high. Studies show that the global prevalence of depression in the perinatal period varies between 10 and 20% [8], and the prevalence of perinatal anxiety disorders varies between 10 and 24% [9]. In addition, the coexistence of depression and anxiety disorder is common in women during the perinatal period [10], and in some studies, this prevalence reaches 40% [11]. However, despite their high prevalence, both disorders are often underdiagnosed [12]. Although it varies according to different periods of pregnancy and working methods in Turkey, the prevalence of pregnancy depression has been determined as 12-36% [13], [14]. Failure to diagnose these common diseases in a timely manner has adverse effects on both mother and baby. In cases that cannot be detected in the early period, it can cause problems in compliance with treatment, functional problems in maternal life, worsening of medical diseases, deterioration of interpersonal relations and economic losses, smoking and substance use, developmental problems, and health problems in the baby. Although anxiety and depression are diseases that are known by physicians and whose negative consequences are better understood every day, they are not sufficiently recognized in the perinatal period, and even if they are recognized, they are not treated quickly enough. Many factors should be considered while determining risk factors in the diagnosis of the disease [15]. These mental illnesses, unplanned pregnancy, unemployed spouse, insufficient social support, low income, low education level, past and present exposure to violence (verbal, physical, sexual), previous pregnancy losses and pregnancy complications, and negative life during pregnancy events can be summarized [16]. In addition, some studies reported risk factors for perinatal anxiety disorder as first birth, multiple births, low-income level, young age, and low social support [17]. In other studies, a history of depressive disorder or an anxiety disorder was determined as the strongest determinant. Considering all these and similar studies, the high number of risk factors causes a loss of time in the evaluation of the patient. In addition, when the study is implemented, it will reach a lot of users, and a lot of data will be obtained. For these reasons, a big data platform is needed to alleviate the workload of women who will give birth, facilitate analysis, and speed up the process. Thus, it is aimed to minimize the potential harm caused by depression and anxiety in women in the perinatal period. In our study, this issue has been discussed, and results that will positively affect the result have been obtained.

To identify pregnant women who were probably experiencing anxiety and/or depression, we used hybrid machine learning techniques that incorporated feature selection methods and classification ML algorithms. Big data architecture was created according to the most optimal solutions obtained from hybrid techniques [18]. The architecture we propose enables us to organize big data analytics in a scalable and efficient manner. In order to prepare for the big data platform, first of all, the data obtained from the patients were cleaned and preprocessed. The ''Feature Selection'' algorithm, which was deemed appropriate for all data, was applied. After this stage, a model that produces the result with acceptable sensitivity was created by using artificial intelligence techniques. A 10-Fold cross-validation technique was used to measure whether the high performance of the model's accuracy is random or not. Among all these artificial intelligence techniques, the machine learning algorithm with the highest performance was used on the big data platform. Thus, a system that can diagnose disease in an instant has been developed. This system has the feature of a platform that serves as an infrastructure for larger data.

With the assistance of this big data platform, the time-consuming procedure of detecting anxiety and depression in pregnant women can be replaced by computer-based systems that operate in real time with considerable accuracy.

The rest of the paper is organized as follows: Section II introduces similar studies related to our study. Section III covers all pre-processing and model development processes, the proposed instant big data platform, and its main components. Section IV presents the experimental results and

discussion. The paper is concluded in Section V with final remarks.

II. LITERATURE REVIEW

In this study; it is aimed to determine the ones that most affect the results of depression and anxiety symptoms and related factors in pregnant women so that the disease can be diagnosed with less data. Then the obtained data are interpreted with the artificial intelligence module and processed in real time on the big data platform. In this context, when we look at the literature, it is seen that various studies have been carried out.

There are a large number of studies using computer technologies in the field of health. The most prominent of these are studies on artificial intelligence. It means interpreting the external data of artificial intelligence systems correctly, learning from such data, and reaching these learnings to certain goals. On the other hand, machine learning is a branch of artificial intelligence in which algorithms discover connections between input and output data [19]. Machine learning algorithms have been used in various branches of healthcare [20], [21]. To calculate the probability of developing acute graft-host illness, Arai et al. used a Japanese population of 26,695 patients. The training group models were developed using the decision tree (ADTree) algorithm. At the end of the applied procedures, the algorithms produced clinically acceptable and highly accurate classification scores. It is very important that the accuracy and precision of the algorithms used are high, especially in the field of health. An ML-based system that identifies cardiac illnesses with 97% accuracy was presented by Yaseen et al. [22]. Another study is the use of neural networks in blood analysis services. When tested on pathological samples, the machine learning system used achieved a diagnostic efficiency of 91% [23]. Machine learning techniques have been used in neurology as well as in other branches of medicine. A useful diagnostic element in neurology is Electroencephalography (EEG) values, which provide insight into the electrical activity of the brain. Many machine learning techniques have been applied to analyze these signals and provide a prediction. Subasi [24] aimed to detect epileptic seizures in EEG recordings by using hybrid SVM (Support Vector Machines) and GA (Genetic Algorithms) algorithms in the EEG dataset. As a result, he obtained a hybrid algorithm providing 99.38% accuracy.

Artificial intelligence techniques are also used in psychological studies. One of them is the work of Priya et al. In this study, it is aimed to detect anxiety and depression by using machine learning algorithms. Priya et al. [25] used five different algorithms (Decision trees, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest) to achieve this. As a result, they concluded that the model that produced the best result was created with the Random Forest algorithm. Another study aimed to detect anxiety and depression in seafarers using machine learning. While doing this, primarily feature selection algorithms were used. In this study conducted among 470 seafarers, five different classification algorithms (Catboost, Random Forest, Logistic Regression, Support Vector Machine (SVM), Naïve Bayes) were used. It has been concluded that the Catboost algorithm is the algorithm that gives the most accurate result, with an accuracy rate of 82.6% [26]. In another study, Paulo et al. used machine learning algorithms to classify major depressive illness (MDD) and bipolar illness (BD) [27]. In order to produce the model with the highest accuracy rate, a 5-fold cross-validation technique was used, and the XGBoost algorithm was preferred as the machine learning algorithm [28]. They also benefited from machine learning techniques at the stage of reaching the optimum solution by adjusting the amounts of therapeutic drugs. One of them is the work of Shuzhe Zhou et al. Major depressive disorder (MDD) patients are often treated with selective serotonin reuptake inhibitors (SSRIs); however, the cure rate is not sufficient. It was intended to develop machine learning models to forecast 8-week outcomes in patients taking SSRIs in order to optimize this [29].

A lot of questions are asked to the patient in diagnosing the psychological disorders that appear in the perinatal period, and this creates a great handicap in answering the questions of the patients. The impatience of patients in answering questions has a negative impact on diagnosis. Choosing the features that affect the disease the most among these questions will both reduce the processing complexity (hence the delay) of artificial intelligence algorithms and reduce the effects of problems such as overfitting and multicollinearity. Some studies are similar to our feature extraction study [30]. Yiye et al. studied the postpartum period data covering one year after birth. In these data, they aimed to make a framework that observes whether women in the postpartum period are depressed or not. They have applied a minimum feature list extraction, processing, and machine learning algorithm from datasets to perform the risk estimation. In training the model with the best accuracy, the dataset was reduced by utilizing clinical features related to mental health history, ultrasonographic complications, drug prescription orders, and patient demographics. After all of this, the model performances, as shown by the area under the receiver operating characteristic curve (AUC) in the development and validation datasets, respectively, are 0.937 and 0.886 [31].

In some instances, the amount of the data may result in performance losses in the operation of the system and resources, regardless of how much the retrieved data is reduced by methods like feature extraction. Especially in the field of health, which is an unprecedented data flow from many sources that contribute to big data in terms of volume, speed, and variability, it is known that instantaneous interpretation of data is of high importance in terms of performance. In such cases where performance and speed are important, big data processing platforms, including machine learning (ML) algorithms, can be used. The vast amount of data that is available in the medical field appears to be manageable by using machine learning approaches to big data on these platforms.

Such methods can provide objective, comparable accuracy measures and significant insight into disease models, opening up possibilities that have not previously been accessible to exploratory medical research. By gathering and examining the existing disease models, big data and machine learning can be employed together in psychiatry [32]. It can support the development of more accurate and appropriate hypotheses for the understanding and treatment of mental diseases. Manal et al. developed a hybrid model to detect chronic lung disease. Effective features were selected using Relief-F and Chi-Squared techniques on the result. Afterward, machine learning algorithms were tested, and performance comparisons were made. All work is done on Apache Spark, a big data processing platform. According to the study, it was observed that the classification algorithms with the best performance were SVM, DT, and GBT, and the selection algorithm with the best performance was Relief-F [18]. Another study in the field of health on big data is that of Lekha et al. In this study, streaming health data was processed, and the patient was informed about their condition in real time by sending a message [1].

It is a fact that the rate of anxiety and depression in women in the perinatal period is relatively high [8], [9], [10], [11]. Therefore, pregnant women in this period are one of the vulnerable groups of the population for mental health disorders. The questions of the dataset created in order to diagnose these people were created by specialist doctors and used for the first time in this study. Different electronic databases of the scientific literature have been extensively searched, but the published articles were found insufficient in screening for anxiety and depression in pregnant women by using machine learning on the big data platform with this dataset. However, there are studies carried out to diagnose different diseases with different datasets [33], [34], [35]. One of them was constructed by Ahmet Husseini et al. In this study, a study on depression and post-traumatic stress disorder was conducted, and the accuracy rate was found to be 80%. In our study, depression and anxiety in women in the perinatal period were studied, and an accuracy rate of 90.80% was obtained using the Naive Bayes algorithm. Since this rate is quite high, it can be said that the system created is very suitable for use in the field of mental health.

Our system also incorporates scalable and high-performance data analytics tools like Apache Kafka and Apache Spark, indicating that this novel architecture can be utilized effectively in instant big data processing applications.

III. MODEL DEVELOPMENT

The proposed system of diagnosing depression and anxiety involves three basic approaches, as seen in Figure 1. The first method selects the crucial features from the perinatal datasets using feature selection methods. The second approach applies Machine learning algorithms such as Random Forest, Decision Tree, Naive Bayes, K-Nearest Neighbors (K-NN), Gradient Boosted Tree (GBT), Logistic Regression, and Deep Feed Forward Neural Network (DFFNN) on the selected features to predict depression and anxiety. In the third approach, in the big data processing platform, the optimum machine learning algorithm was utilized on all the data, which was cleaned and made ready by using feature selection algorithms.

In this context, the suggested method consists of seven processes, the first of which is data collecting using the perinatal dataset from real patients. The handling of null values, identification of outlier values, and data integration are made in the second stage (data pre-processing step). The third stage is involved choosing the important features using feature selection methods. The data set is divided into testing/training in the ML process in the fourth stage. The cross-validation method was used to optimize this process. In the fifth stage, the machine learning that gives the performance of prediction in the most optimal way is decided among seven machine learning algorithms.

All the data, which is cleaned and made ready by using feature selection algorithms, is sent to Apache Spark, which is the Consumer from Kafka Producer, at certain second intervals at the sixth stage. At the seventh stage, the Consumer Apache Spark has received the stream in a highly efficient, fault-tolerant manner thanks to the Streaming API, and the incoming data has interacted with the Naive Bayes algorithm, which has the highest performance, through the Apache Spark MLLIB API, and the disease results are output. All processes are shown in Figure 2.

For the operation of these seven processes, the open-source RapidMiner (RapidMiner Studio 9.10.10) [36] and Apache Spark (version 3.0.1) [2], a big data processing platform, were used. Java was selected as the preferred programming language for Apache Spark.

The following subsections contain a detailed explanation of each stage.

A. DATA COLLECTION

Data were collected from pregnant women who applied to Sakarya University Training and Research Hospital's obstetrics outpatient clinic. Three hundred ninety-five people had to be asked before reaching the target of two hundred fifty women who agreed to participate in the study and fill out the forms. Pregnant women were asked to fill in SDVF (Socio-demographic Data Form), PASS-TR (Perinatal Anxiety Symptom Scale) [37], and EPDS (Edinburgh Postpartum Depression Scale) [38] tests. The SDVF consists of questions created by specialist doctors for this study. PASS-TR and EPDS are scales that detect depression and anxiety, respectively. A patient is classified as having anxiety or depression based on the ''label'' column used for labeling. While creating the dataset, EPDS and PASS outputs are taken as label values, respectively. The level of depression was assessed using the EPDS scale.

Score less than 13 was considered as not significant depression. A score lower than 13 was regarded as not significant depression (labeled with a ''0'' meaning ''No

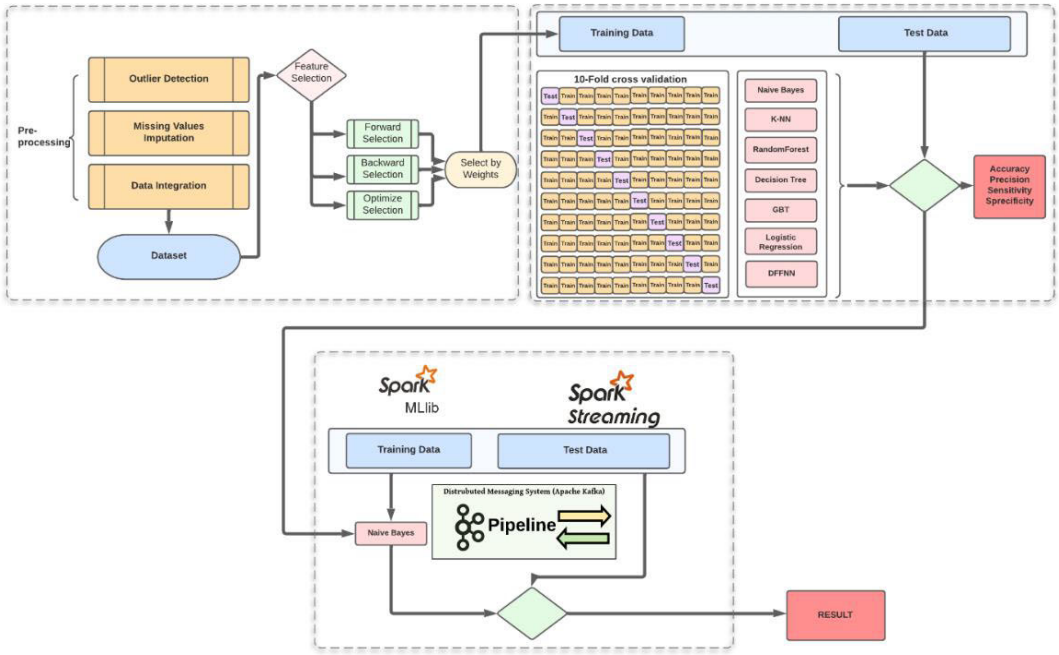

FIGURE 1. Outline of the proposed data analytics architecture.

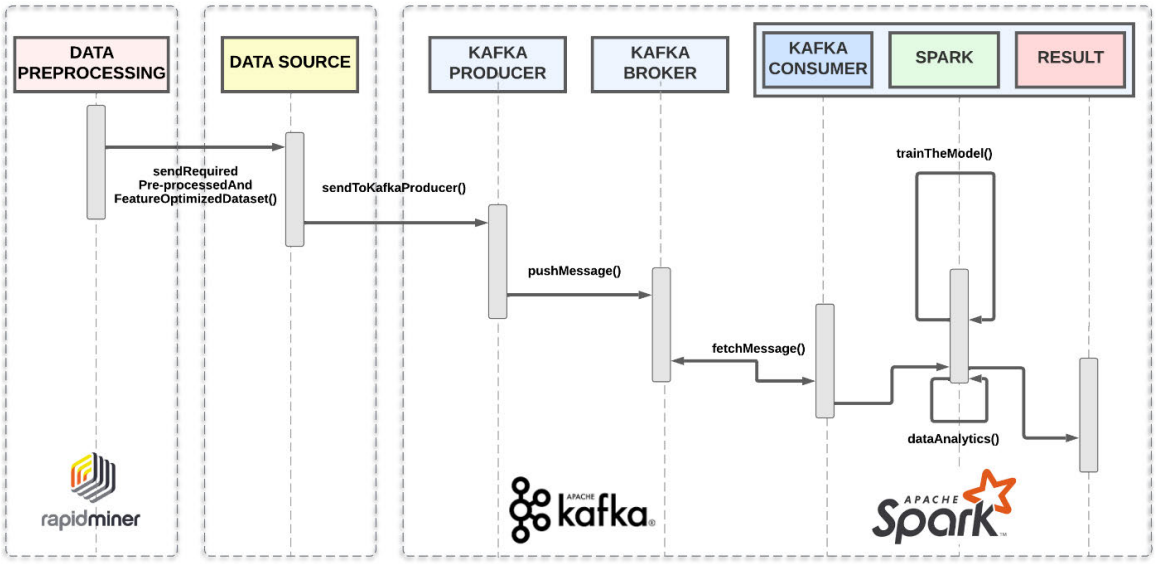

FIGURE 2. The sequence diagram for end-to-end data delivery operations.

Depression''), otherwise considered as depression (labeled with a ''1'' meaning ''Depression''). The PASS scale was used to measure the degree of anxiety. Score less than 17 was considered as not significant anxiety (labeled with a ''0'' meaning ''No Anxiety''); otherwise suffering from significant anxiety disorder (labeled with a ''1'' meaning ''Anxiety''). A person can be classified as ''No Anxiety-No Depression,'' ''No Anxiety-Depression,'' ''Anxiety-No Depression,'' or ''Anxiety-Depression'' based on their EPDS and PASS scores. All groups, with the exception of ''No Anxiety-No Depression,'' require treatment for mental health disorders by a psychiatrist. We decided to group all the outputs into two labels: ''No Anxiety-No Depression (0)'' and ''Anxiety and/or Depression so (1)'' since the purpose of this research is to evaluate the applicability and suitability of the machine learning technology of big data platforms (Table 1.)

TABLE 1. Classification label.

Anxiety (PASS)	Depression (EPDS)	Label
0	0	0
1	0	1
0	1	1
1	1	1

"Anxiety and/or Depression" members of the group are to be referred to a psychiatrist for a diagnostic assessment and appropriate management since they may be experiencing anxiety, depression, or both [26].

The perinatal pregnancy dataset includes 250 samples, 60 features, and 1 class label. The class label has two values: "No Anxiety-No Depression (0)" and "Anxiety and/or Depression (1)". The details of each feature are described below.

Features used as predictive variables included as follows;

- socio-demographic information (SDVF) regarding age, marital status, level of physical activity, years of education, working status, total income level of the household, psychological history, psychological traumas caused by her first pregnancy, the psychological history of her husband and family
- anxiety levels measured through the PASS, which assesses the existence and severity of present anxiety symptoms
- depression symptoms' severity was evaluated with Edinburgh Postnatal Depression Scale (EPDS)

All data were used for the diagnosis of anxiety and depression. All clinical assessments were performed by experienced and trained psychiatrists on study baselines.

B. DATA PRE-PROCESSING

The dataset obtained from the patients included missing values and outliers. Then, it needs to be cleaned and unblemished during the pre-processing stage. The missing value estimation, normalization, unbalanced data checking, and outliers have all been part of the pre-processing stage [18]. The simplest way to deal with missing values is to ignore the dataset. Rather than deleting records, we filled in what could be filled in the missing values. Averaging has been used to fill in the missing values of nominal characteristics. In order to give more meaningful results in our analysis, the data were grouped and then categorized. Outliers were identified using the local outlier factor (LOF) approach, which calculates outliers based on density. The LOF approach is a computational method used to identify local outliers by providing outliers on a numerical scale, such as the extent to which an object is isolated from its surrounding neighbors. As a result, clean data was obtained after these processes.

C. FEATURE SELECTION METHODS

The primary advantage of using feature selection algorithms is identifying the dataset's important features. The classifier approach with feature selection yields better results while significantly reducing model execution time [39]. In this study, the patient is asked a lot of questions that need to be answered in order to reach the result. This causes a waste of time and reduces the patient's continuity during the treatment. But even so, it is important to specify which questions of the SDVF will be used. We trained our algorithms with various combinations of the applied form in order to analyze this effect. The following combinations were tested in this study: forward selection, backward elimination, and optimized selection-evolutionary trait selection. The optimized selection-evolutionary technique showed the greatest effect on model performance. In this technique, the genetic algorithm mimics the course of natural evolution. According to the Feature Selection algorithms made on the SDVF dataset, the features with the highest priority are in Figure 3. As can be seen from this figure, the most important features out of 60 features are listed according to their weights from the most to the least. The equivalents and weights of these features in SDVF form are shown in Table 2.

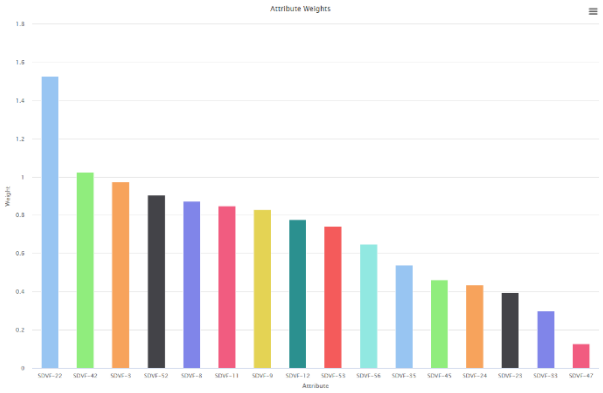

FIGURE 3. The important features and their weights selected by the feature selection algorithm.

According to the outputs, it is possible to say that the most important feature is the "gestational week" with a weight of 1.527, and the least important feature among the most important features is "whether or not her husband has a chronic illness" with a weight of 0.127.

D. SPLITTING THE DATASET-CROSS VALIDATION

Cross Validation is an algorithm used in model selection to better predict the error of a test performed on a machine learning model. In our study, a 10-fold cross-validation algorithm was used to separate the model as training and test data. The 10-fold cross-validation algorithm divides the data set into ten equal parts. One of these pieces is used as the test dataset, and the remaining nine pieces are used as the training dataset. This process is repeated ten times until each piece is a test data set. The accuracy of the model is determined by averaging the accuracy values.

E. MODELING

The following machine learning algorithms were used in the study: Decision Tree, Naive Bayes, K-NN, Random Forest,

TABLE 2. The most important features correspond to the dataset.

Features	Name in dataset	Weight of Features
Gestational week	SDVF-22	1.527
Baby's health problem	SDVF-42	1.025
Feature of the living environment (urban or rural)	SDVF-3	0.976
Communication with her husband	SDVF-52	0.905
Her level of education	SDVF-8	0.872
Number of people living in the house	SDVF-11	0.849
Working status	SDVF-9	0.830
Total income level	SDVF-12	0.778
Emotional support of her husband	SDVF-53	0.742
Presence of people to share her problems with	SDVF-56	0.649
Exercise status	SDVF-35	0.540
Her husband's educational status	SDVF-45	0.463
Whether there is a desired pregnancy	SDVF-24	0.436
Baby's gender	SDVF-23	0.394
Smoking	SDVF-33	0.298
Her husband chronic illness	SDVF-47	0.127

GBT, Logistic Regression, and DFFNN. As was shown in Table 3., since the Naive Bayes algorithm gave the highest performance, the use of the Naive Bayes algorithm was deemed appropriate in the continuation of the study.

Naive Bayes (NB): The Bayes theorem is used to train a classifier for the Naive Bayes approach. It is predicated on the idea that one feature in a decision class does not exclude the existence of another feature in that class.

Assume that C = {c1, c2, …, cm} is a set of classes and that X = {x1, x2,…, xn} is a set of features. Equation (5) can be used to determine the posterior probability for each class variable given a given characteristic from the Bayes theorem.

$$P(C_i|X_J) = \frac{P(X_J|C_i)P(C_i)}{P(X_J)} \tag{1}$$

where, i = 1,2, …m and j = 1,2, …, n, $P(X_J|C_i)$ = Probability of feature X_j with given class C_i, $P(X_J)$ = Prior probability of feature X_j and $P(C_i)$ = Prior probability of class C_i. The classifier then determines the class variables' highest probability. The classification result will be the class with the highest posterior probability value [26].

F. EVALUATING THE MODELS

The models are assessed using five common metrics, as stated in Equations 2-6: accuracy, sensitivity, specificity, precision, and area under the curve (AUC) of ROC, where TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative [18].

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

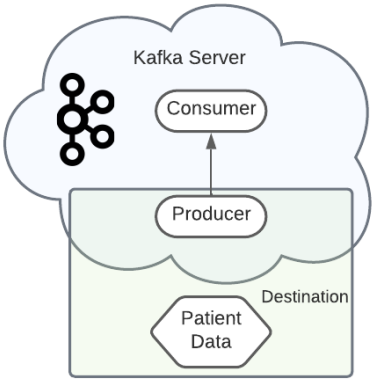

FIGURE 4. Streaming destination and Kafka (producer/consumer).

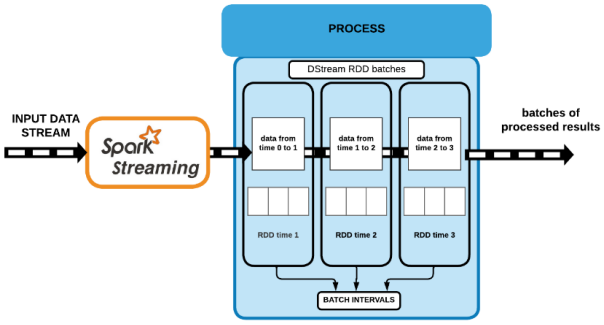

FIGURE 5. The general structure of the spark streaming process [2].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{AUC} = \textit{Area under the curve of ROCi .e. TP vs FP} \tag{6}$$

G. APACHE KAFKA

Apache Kafka (version 3.0.2) [5] is a high-throughput message-distribution system that can manage enormous volumes of data while still providing service to numerous users and producers. Kafka executes parallelism by utilizing numerous partitions and different brokers, which speeds up the transmission and offers unbroken services in the event that a broker fails. Additionally, it has the ability to support streaming real-time data [40].

The Kafka Producer gets cleaned patient data and sends it to the Kafka Broker so that it can be delivered to consumers.

H. APACHE SPARK

Apache Spark (version 3.0.1) [2] is an open-source platform for data processing that can swiftly conduct operations on very large data sets and distribute operations over numerous machines. It utilizes optimized query execution

and in-memory caching for quick analytic queries against any size of data. Spark can be deployed in a variety of ways, provides native bindings for the Java, Scala, Python, and R programming languages, and supports SQL, streaming data, machine learning, and graph processing [40].

1) APACHE STREAMING

Apache Spark (version 3.0.1) [2] is an open-source platform for data processing that can swiftly conduct operations on very large data sets and distribute operations over numerous machines. It utilizes optimized query execution and in-memory caching for quick analytic queries against any size of data. Spark can be deployed in a variety of ways, provides native bindings for the Java, Scala, Python, and R programming languages, and supports SQL, streaming data, machine learning, and graph processing [40].

2) APACHE MLLIB

Apache Spark has a scalable library called MLlib that contains typical machine learning algorithms such as regression, classification, clustering, pattern mining, and collaborative filtering. The naive bayes algorithm, which gives the highest accuracy compared to other algorithms, was selected in this study to classify patient data. The multi-class classification algorithm named Naive Bayes makes the proposition that each feature pair is independent of the other. It applies Bayes' theorem to the training data in a single pass and then uses that information to predict the conditional probability distribution of a label given an observation [2].

IV. EXPERIMENTAL RESULTS

In this study, datasets were created using clinical variables collected through self-administered questionnaires. Clinical variables include EPDS, PASS scales, and SDVF. This dataset investigates whether patients have anxiety, depression, or both. This study is aimed to minimize the potential harms of depression and anxiety in women. The results obtained in this study are gathered under the headings of classification and streaming results.

A. CLASSIFICATION RESULTS

Seven different machine learning algorithms (Decision Tree, Naive Bayes, K-NN, Random Forest, GBT, Logistic Regression, and DFFNN) were applied to the patient data consisting of the features shown in Figure 3. Performance values are shown in Table 3.

The performance of cross-validation of applying ML to the features selected Optimize Selection has achieved the best by Naive Bayes with 90.80% accuracy, 86.90% sensitivity, 92.34% specificity, 81.71% precision, and 0.966 AUC as shown in Table 3. According to the results shown in Table 3, the machine learning algorithm with the highest performance, Naive Bayes, was then used to train the machine learning model on the big data processing platform. Results from this platform are in streaming results section has also been given.

TABLE 3. Machine learning model performance metrics.

Classifier	Accuracy	Sensitivity	Specificity	Precision	AUC
Decision Tree	87.20%	74.05%	92.22%	80.05%	0.822
Naive Bayes	90.80%	86.90%	92.34%	81.71%	0.966
K-Nearest Neighbor	86.80%	68.57%	93.89%	81.95%	0.900
Random Forest	88.00%	69.29%	95.00%	86.10%	0.945
Gradient Boosted Tree	87.60%	77.14%	91.70%	77.14%	0.938
Logistic Regression	84.80%	73.81%	88.92%	74.58%	0.908
Deep Feed Forward Neural Network	89.60%	80.00%	93.33%	83.26%	0.962

TABLE 4. Naive bayes model's evaluated parameters.

Parameter	Value
Accuracy	92.45%
Precision	97.29%
Sensitivity	92.30%
F1	92.60%

B. STREAMING RESULTS

All patient data were pre-processed and prepared using the optimum feature selection algorithm in the previous stages. This data pushes into Kafka Broker at a certain number of seconds (1 second) by Kafka Producer. As soon as the Kafka Producer pushes this data into the Kafka Broker, Kafka Consumer fetches it for use in its respective jobs. The Kafka Consumers are connected to the Spark engine, as shown in Figure 2. When Apache Spark receives patient-related patient data, it instantly diagnoses disease using the Naive Bayes model.

Apache Spark has an MLlib module that supports common machine learning algorithms. The naive bayes algorithm was selected in this study to classify patient data because it gives higher performance than other classifications shown in Table 3. Evaluation metrics of the Naive Bayes algorithm used in Apache Spark are shown in Table 4.

80% of the patient dataset was used to train and test the developed Naive Bayes model. This model interacted with streaming data generated by 20% of the dataset.

Fig. 6 displays incoming test data, as well as the naive bayes algorithm classification of the data in an instant and the predicted results according to the data training model. Patient data are the socio-demographic data shown in table 2, which are most influential on the result of the feature selection

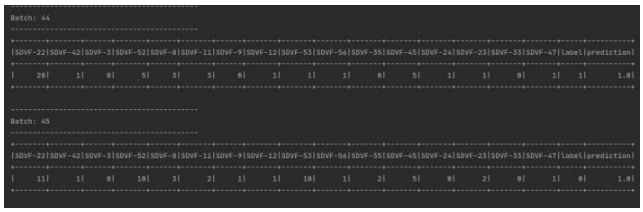

FIGURE 6. Streaming result predicted by naive bayes.

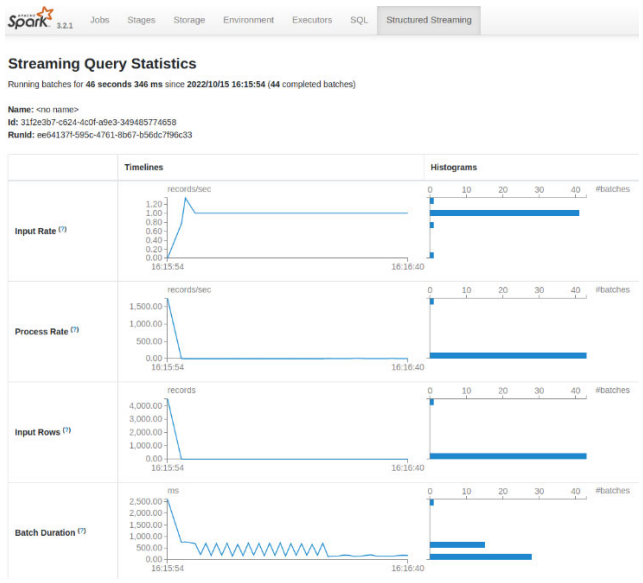

FIGURE 7. High input rate in 1-second data streaming flow.

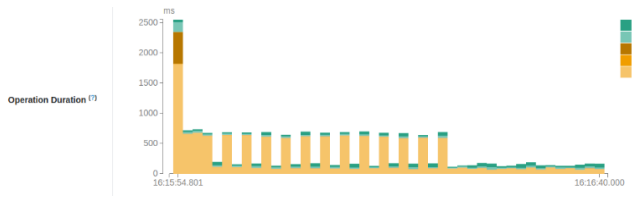

FIGURE 8. Operation duration of streaming flow.

algorithms. The ''label'' column used in labeling is the class rating. The label column has two values: ''No Anxiety-No Depression (0)'' and ''Anxiety and/or Depression so (1).

The result of the two-second streaming data, which provides patient data every second, is shown in Figure 6.

The two patients' most important features were analyzed at one-second intervals, and the findings were shown. For example, for batch number 44 in the first row, which shows the value of SDVF-22 as 20 and the value of SDVF-42 as 1, the label rating value is equal to the expected value of the naive bayes. Here, the naive bayes classifies the record as positive for the disease. The label value is not equal to the expected naive bayes at batch number 45 in the second row, which shows the value of SDVF-22 as 11 and SDVF-42 as 1. Here, the naive bayes model incorrectly classifies this record as positive for the disease.

Stream processing requires continuous data stream handling within a short period (from a few milliseconds to minutes). The results were obtained using one Spark executor that handles streaming data. Figure 7 and figure 8 illustrates the Spark web user interface (WebUI), which displays the Spark Streaming performance results, including input rate, process rate, input rows, batch duration, and operation duration. They are valuable results for examining data processing capability in real-time [40].

The input rate aggregates (across all sources) the rate of data arriving. The process rate shows how quickly data is analyzed. The input rows show the total number of records handled by a trigger. The batch duration demonstrates the process duration of each batch. Operation duration means the amount of time taken to perform various operations in milliseconds [2].

V. CONCLUSION

There are many times in people's lives that will negatively affect their psychological balance. One of the most crucial periods affecting mental health is during the perinatal period. If adequate treatment is not applied in this time period, it is an inevitable fact that it will cause serious negativities on the child, the mother, and the mental health of society afterward. It is an undeniable fact that in such a vital area, there is a need for a system that enables the patient to reach diagnosis and treatment faster and easier, but there is no study aiming to meet women with treatment through a hybrid big data analytics platform as instantly. This study concentrates, in particular, on hybrid big data platforms that use machine learning techniques to detect mental health conditions.

In order to reduce the questions of form and scales that women have to answer and to identify pregnant women who were likely suffering anxiety and/or depression more quickly, the data that has more importance on the result of the dataset was selected with feature selection algorithms. Sequential Forward (SFS), Sequential Backward (SBS), and Optimized feature selection techniques are used. It has been observed that the optimized selection technique in which genetic algorithms are used is the technique with the highest performance. Six different machine learning classifiers (Decision Tree, Naive Bayes, K-NN, Random Forest, GBT, Logistic Regression, and DFFNN) were evaluated for their effectiveness and efficacy. The ML parameters were optimized using cross-validation.

Six evaluation methods, accuracy, sensitivity, specificity, and precision, were applied to validate the results, and the testing data were registered. The results showed that the Naive Bayes Classifier with the selected features had achieved the best performance with 90.80% accuracy. Thanks to this high accuracy rate, it is not difficult to say that the model we developed can be used effectively for diagnosis.

The Naive Bayes-based machine learning model was built on the Apache Spark platform. This model interacted with streaming data sent via Apache Kafka to Apache Spark.

Thus, the developed system was tested for instant detection of anxiety/depression in pregnant women.

This research emphasizes hybrid big data platforms using machine learning technology in the field of the detection of mental health diseases. It is obvious that the suggested architecture may be utilized successfully in instant big data processing applications since it incorporates scalable and high-performance data analytics tools like Apache Kafka and Spark. With the use of this architecture, the time-consuming analysis of anxiety and depression can be replaced by an automated computer-based technique with a reasonable amount of accuracy at the moment.

ETHICAL STATEMENT

Informed consent was obtained from every study participant before being interviewed. Ethics committee approval was taken from Sakarya University Research Hospital for the data used in this article. We confirm that ethical approval has been granted, and the reference number is E-71522473-050.01.04-112781-38.

ACKNOWLEDGMENT

The authors would like to thank the consultant psychiatrists at the Department of Psychiatry for their valuable suggestions and expert validation of the interview questionnaire.

DATA AVAILABILITY

Data were collected from pregnant women who applied to Sakarya University Training and Research Hospital's obstetrics outpatient clinic for just this study.

REFERENCES

[1] L. R. Nair, S. D. Shetty, and S. D. Shetty, "Applying spark based machine learning model on streaming big data for health status prediction," *Comput. Electr. Eng.*, vol. 65, pp. 393–399, Jan. 2018, doi: 10.1016/j.compeleceng.2017.03.009.

[2] Apache Spark. (2018). *Apache Spark™—Unified Analytics Engine for Big Data*. [Online]. Available: https://spark.apache.org

[3] *Apache Hadoop*. Accessed: Oct. 26, 2022. [Online]. Available: https://hadoop.apache.org/

[4] *Apache Storm*. Accessed: Oct. 26, 2022. [Online]. Available: https://storm.apache.org/

[5] *Apache Kafka*. Accessed: Oct. 27, 2022. [Online]. Available: https://kafka.apache.org/

[6] A. M. Y. Tai, A. Albuquerque, N. E. Carmona, M. Subramanieapillai, D. S. Cha, M. Sheko, Y. Lee, R. Mansur, and R. S. McIntyre, "Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry," *Artif. Intell. Med.*, vol. 99, Aug. 2019, Art. no. 101704, doi: 10.1016/j.artmed.2019.101704.

[7] D. Bzdok and A. Meyer-Lindenberg, "Machine learning for precision psychiatry: Opportunities and challenges," *Biol. Psychiatry, Cogn. Neurosci. Neuroimag.*, vol. 3, no. 3, pp. 223–230, Mar. 2018, doi: 10.1016/j.bpsc.2017.11.007.

[8] N. B. Dorio, S. S. Fredrick, and M. K. Demaray, "School engagement and the role of peer victimization, depressive symptoms, and rumination," *J. Early Adolescence*, vol. 39, no. 7, pp. 962–992, Aug. 2018, doi: 10.1177/0272431618797007.

[9] L. M. Howard and H. Khalifeh, "Perinatal mental health: A review of progress and challenges," *World Psychiatry*, vol. 19, no. 3, pp. 313–327, Oct. 2020, doi: 10.1002/WPS.20769.

[10] C.-L. Dennis, K. Falah-Hassani, and R. Shiri, "Prevalence of antenatal and postnatal anxiety: Systematic review and meta-analysis," *Brit. J. Psychiatry*, vol. 210, no. 5, pp. 315–323, May 2017, doi: 10.1192/BJP.BP.116.187179.

[11] C. Reck, K. Struben, M. Backenstrass, U. Stefenelli, K. Reinig, T. Fuchs, C. Sohn, and C. Mundt, "Prevalence, onset and comorbidity of postpartum anxiety and depressive disorders," *Acta Psychiatrica Scandinavica*, vol. 118, no. 6, pp. 459–468, Dec. 2008, doi: 10.1111/j.1600-0447.2008.01264.x.

[12] S. Misri, J. Abizadeh, S. Sanders, and E. Swift, "Perinatal generalized anxiety disorder: Assessment and treatment," *J. Women's Health*, vol. 24, no. 9, pp. 762–770, Sep. 2015, doi: 10.1089/JWH.2014.5150.

[13] D. Aslan and R. Edelmann, "Demographic and offence characteristics: A comparison of sex offenders convicted of possessing indecent images of children, committing contact sex offences or both offences," *J. Forensic Psychiatry Psychol.*, vol. 25, no. 2, pp. 121–134, Mar. 2014, doi: 10.1080/14789949.2014.884618.

[14] T. S. Kirkan, N. Aydin, E. Yazici, P. Akcali Aslan, H. Acemoglu, and A. G. Daloglu, "The depression in women in pregnancy and postpartum period: A follow-up study," *Int. J. Social Psychiatry*, vol. 61, no. 4, pp. 343–349, Jun. 2015, doi: 10.1177/0020764014543713.

[15] D. Hubbeling, "Overtreatment: Is a solution possible?" *J. Eval. Clin. Pract.*, vol. 28, no. 5, pp. 821–827, Oct. 2022, doi: 10.1111/jep.13632.

[16] M. Furtado, C. H. T. Chow, S. Owais, B. N. Frey, and R. J. Van Lieshout, "Risk factors of new onset anxiety and anxiety exacerbation in the perinatal period: A systematic review and meta-analysis," *J. Affect. Disorders*, vol. 238, pp. 626–635, Oct. 2018, doi: 10.1016/J.JAD.2018.05.073.

[17] J. H. Goodman, K. L. Chenausky, and M. P. Freeman, "Anxiety disorders during pregnancy: A systematic review," *J. Clin. Psychiatry*, vol. 75, no. 10, p. 1177, Oct. 2014, doi: 10.4088/JCP.14R09035.

[18] M. A. Abdel-Fattah, N. A. Othman, and N. Goher, "Predicting chronic kidney disease using hybrid machine learning based on apache spark," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Feb. 2022, doi: 10.1155/2022/9898831.

[19] C. L. Bormann, M. K. Kanakasabapathy, P. Thirumalaraju, R. Gupta, R. Pooniwala, H. Kandula, E. Hariton, I. Souter, I. Dimitriadis, L. B. Ramirez, C. L. Curchoe, J. Swain, L. M. Boehnlein, and H. Shafiee, "Performance of a deep learning based neural network in the selection of human blastocysts for implantation," *eLife*, vol. 9, pp. 1–14, Sep. 2020, doi: 10.7554/ELIFE.55301.

[20] Y. Arai, T. Kondo, K. Fuse, Y. Shibasaki, M. Masuko, J. Sugita, T. Teshima, N. Uchida, T. Fukuda, K. Ohashi, Y. Ozawa, T. Ichinohe, Y. Kanda, and Y. Atsuta, "Prediction of acute graft-versus-host disease following allogeneic hematopoietic stem cell transplantation using a machine learning algorithm," *Blood*, vol. 132, p. 68, Nov. 2018, doi: 10.1182/BLOOD-2018-99-114794.

[21] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification prediction of breast cancer based on machine learning," *Comput. Intell. Neurosci.*, vol. 2023, Jan. 2023, Art. no. 6530719.

[22] Yaseen, G.-Y. Son, and S. Kwon, "Classification of heart sound signal using multiple features," *Appl. Sci.*, vol. 8, no. 12, p. 2344, Nov. 2018, doi: 10.3390/APP8122344.

[23] G. Zini, "Artificial intelligence in hematology," *Hematology*, vol. 10, no. 5, pp. 393–400, Oct. 2005, doi: 10.1080/10245330410001727055.

[24] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 1084–1093, May 2007, doi: 10.1016/J.ESWA.2006.02.005.

[25] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Proc. Comput. Sci.*, vol. 167, pp. 1258–1267, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.442.

[26] A. Sau and I. Bhakta, "Screening of anxiety and depression among the seafarers using machine learning technology," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100149, doi: 10.1016/J.IMU.2018.12.004.

[27] P. J. C. Suen, S. Goerigk, L. B. Razza, F. Padberg, I. C. Passos, and A. R. Brunoni, "Classification of unipolar and bipolar depression using machine learning techniques," *Psychiatry Res.*, vol. 295, Jan. 2021, Art. no. 113624, doi: 10.1016/J.PSYCHRES.2020.113624.

[28] Y. Luo, Z. Wang, and C. Wang, "Improvement of APACHE II score system for disease severity based on XGBoost algorithm," *BMC Med. Informat. Decis. Making*, vol. 21, no. 1, pp. 1–12, Dec. 2021, doi: 10.1186/S12911-021-01591-X.

[29] S. Zhou, Q. Ma, Y. Lou, X. Lv, H. Tian, J. Wei, K. Zhang, G. Zhu, Q. Chen, T. Si, G. Wang, X. Wang, N. Zhang, Y. Huang, Q. Liu, and X. Yu, "Machine learning to predict clinical remission in depressed patients after acute phase selective serotonin reuptake inhibitor treatment," *J. Affect. Disorders*, vol. 287, pp. 372–379, May 2021, doi: 10.1016/J.JAD.2021.03.079.

[30] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Comput. Biol. Med.*, vol. 112, Sep. 2019, Art. no. 103375, doi: 10.1016/J.COMPBIOMED.2019.103375.
[31] Y. Zhang, S. Wang, A. Hermann, R. Joly, and J. Pathak, "Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women," *J. Affect. Disorders*, vol. 279, pp. 1–8, Jan. 2021, doi: 10.1016/J.JAD.2020.09.113.
[32] H. Cao, A. Meyer-Lindenberg, and E. Schwarz, "Comparative evaluation of machine learning strategies for analyzing big data in psychiatry," *Int. J. Mol. Sci.*, vol. 19, no. 11, p. 3387, Oct. 2018, doi: 10.3390/IJMS19113387.
[33] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol., From Keyboard Clinic*, 2018, pp. 88–97.
[34] R. C. Kessler, H. M. van Loo, K. J. Wardenaar, R. M. Bossarte, L. A. Brenner, T. Cai, D. D. Ebert, I. Hwang, J. Li, P. de Jonge, A. A. Nierenberg, M. V. Petukhova, A. J. Rosellini, N. A. Sampson, R. A. Schoevers, M. A. Wilcox, and A. M. Zaslavsky, "Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports," *Mol. Psychiatry*, vol. 21, no. 10, pp. 1366–1371, Oct. 2016, doi: 10.1038/MP.2015.198.
[35] P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, and D. S. Lee, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes," *J. Clin. Epidemiol.*, vol. 66, no. 4, pp. 398–407, Apr. 2013, doi: 10.1016/J.JCLINEPI.2012.11.008.
[36] *RapidMiner | Amplify the Impact of Your People, Expertise & Data*. Accessed: Oct. 27, 2022. [Online]. Available: https://rapidminer.com/
[37] E. Yazıcı, T. M. Pek, H. U. Yuvacı, E. Köse, S. Cevrioglu, A. B. Yazıcı, A. S. Çilli, A. Erol, and N. Aydın, "Perinatal anxiety screening scale validiy and reliability study in Turkish (PASS-TR validity and reliability)," *Psychiatry Clin. Psychopharmacol.*, vol. 29, no. 4, pp. 609–617, Oct. 2019, doi: 10.1080/24750573.2018.1506247.
[38] J. L. Cox, J. M. Holden, and R. Sagovsky, "Detection of postnatal depression: Development of the 10-item Edinburgh postnatal depression scale," *Brit. J. Psychiatry*, vol. 150, no. 6, pp. 782–786, Jun. 1987, doi: 10.1192/BJP.150.6.782.
[39] Y. Li, D.-I. Stroe, Y. Cheng, H. Sheng, X. Sui, and R. Teodorescu, "On the feature selection for battery state of health estimation based on charging–discharging profiles," *J. Energy Storage*, vol. 33, Jan. 2021, Art. no. 102122, doi: 10.1016/j.est.2020.102122.
[40] N. B. Oğur, M. Al-Hubaishi, and C. Çeken, "IoT data analytics architecture for smart healthcare using RFID and WSN," *ETRI J.*, vol. 44, no. 1, pp. 135–146, Feb. 2022, doi: 10.4218/ETRIJ.2020-0036.

NUR BANU OĞUR received the B.S. and M.S. degrees in computer engineering from Sakarya University, Turkey, in 2015 and 2018, respectively, where she is currently pursuing the Ph.D. degree in computer and information engineering. Since 2017, she has been a Research Assistant with the Department of Computer Engineering, Sakarya University. Her research interests include data analytics, the Internet of Medical Things, and artificial intelligence.

CELAL ÇEKEN received the Ph.D. degree from Kocaeli University, Kocaeli, Turkey, in 2004. Since 2012, he has been with the Computer Engineering Department, Faculty of Computer and Information Sciences, Sakarya University, Sakarya, Turkey. His research interests include wireless communications, wireless sensor/actuator networks, the Internet of Things, and software-defined networking.

YAVUZ SELİM OĞUR received the degree from the Faculty of Medicine, Istanbul University, in 2016. He was a Research Assistant with the Physiology Department, Sakarya University, from 2018 to 2022. He continued his professional life as a Specialist Doctor with the Sakarya Training and Research Hospital after completing the specialization education.

HİLAL USLU YUVACI received the degree from the Faculty of Medicine, Ankara University, in 2001. She was a Research Assistant with the Gynecology and Obstetrics Department, Fatih University. She continued her professional life as a specialist after completing the specialization education. She has been with Sakarya University Training and Research Hospital, since 2017.

AHMET BÜLENT YAZICI was a Specialist Doctor from 2005 to 2017. In 2017, he was an Assistant Professor with the Department of Psychiatry, Faculty of Medicine, Sakarya University, and an Associate Professor, in 2018.

ESRA YAZICI received the degree from the Faculty of Medicine, Atatürk University, in 2003. She was a Research Assistant with the Physiology Department, Atatürk University, from 2004 to 2005, and the Psychiatry Department, Atatürk University, from 2005 to 2010. She continued her professional life as a Specialist Doctor with the Kocaeli Derince Training and Research Hospital after completing the specialization education. In 2014, she was an Assistant Professor with the Department of Psychiatry, Faculty of Medicine, Sakarya University, and an Associate Professor, in 2015, where she has been a Professor Doctor, since 2021.