A Real-time Health Monitoring and Prediction System using Incremental Learning

Appasani Deepthi
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham
Amritapuri, India
deepthiappasani@gmail.com

Sanketh Yelamanchili
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham
Amritapuri, India
sankethy07@gmail.com

Rahul Rishi Pentakota
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham
Amritapuri, India
rahulrishi.pentakota@gmail.com

Charan Sai Bokkisam
Department of Computer Science and Engineering
Amrita Vishwa Vidyapeetham
Amritapuri,India
charansaibokkisam@gmail.com

***Abstract*—Since the dawn of modern computers, scientists and medics have been fascinated by the propensity of the AI approach in medical and healthcare applications. With regard to health care, any information on a patient's or population's health is considered to be health information and this leads to some significant issues such as processing enormous volumes of data that are produced at fast rates and also there is no uniform criteria for the data that devices gather, which might result in data creation and prevent broad adoption. In addition to that, establishing data preservation, cleaning, and retention may be challenging given the enormous volume of data. Since much of the data will be continuous and real-time, typical batch ML methods are unable to solve the aforementioned problems. Here comes a new key term named Incremental Learning to our rescue, which describes the problem of continual model adaption based on an ongoing data stream which fits the ideology of health care analytics of using both recent and old data to generate macro and micro insights that support corporate and patient decision-making using less resources .Incremental methods are truly incremental in the sense that they learn from each training example as it arrives. In this paper, we implemented several incremental learning algorithms to analyse healthcare data that is publicly available. and also implemented several batch algorithms and a variety of preprocessing methods to increase the accuracy of the predictions on the datasets and also did a comparative study based on their performance among them.**

***Index Terms*—Batch learning, Incremental learning, Machine learning, real-time data, Healthcare,IoT**

I. Introduction

Data analytics is the process of examining unprocessed data sets for patterns, conclusions, and prospects for improvement This analytics uses both recent and historical data to generate macro and micro insights that support corporate and patient decision-making. However, there are some significant issues, A system that entails improving medical services in order to meet peoples' medical needs falls under the broad category of health care. Patients, doctors, vendors, health organizations, and IT firms all work to protect and restore patient records in the healthcare industry. Machine learning is being used in healthcare analyses to treat a variety of ailments, including cancer, diabetes, strokes, and other conditions.Machine learning (ML) in healthcare research has been growing substantially over the past few years[8]

A classic batch-learning method is trained on batches of the data in the batch-incremental approach: every w new instances constitute a batch, and once that batch is finished, it is assigned to a learner to train on. These approaches' key drawbacks include the need for a parameter w that specifies batch size, the requirement to discard learned models in order to create room for new ones, and the inability to learn from the most recent examples until a new batch is complete.Both academics and business have recently begun to pay more and more attention to incremental learning. From the perspective of computational intelligence, incremental learning is important for at least two main reasons. First, from a data mining perspective, a lot of today's data-intensive computing applications call for the learning algorithm to be able to incrementally learn from large-scale dynamic stream data and to gradually build up the knowledge base to support subsequent learning and decision-making processes.[3].As long as there are fresh data examples, the learning process continues in incremental learning. According to the knowledge gained from the most recent data instances observed by the algorithm, the existing data model is updated.

Since the system receives data in stages and it is impractical to build a new model each time new data is received, incremental learning techniques are very useful for real-world applications. A data model is originally created and modified when new data becomes available. This approach upgrades the existing model to an accurate one as opposed to conventional methods, which involve building a new model from scratch. [2]. Two major obstacles, concept drift and class imbalance, Concept drift is a colloquial term for a change in the definitions of the classes (concepts) over time, which results in a change in the distributions from which the data for these concepts are derived. [1] [5].The more trustworthy data-driven predictions and decision tools has emerged as a critical challenge in an ever-changing and large data world. [4] In this paper we applied Incremental learning algorithms on healthcare data sets.We implemented batch algorithms along with incremental learning algorithms with several Preprocessing techniques to find the impact of Incremental algorithms on the data.

II. Data Description

The gathering of data is a crucial component of research. These days, healthcare is utilizing a variety of data kinds. Different mining techniques are used on this type of data to extract the more pertinent features, and then various algorithms must be trained for improved future prediction. Three data sets are considered for the prediction analysis in this research.[8]

- Hepatocellular Carcinoma data set (HCC data set)
- Fetal Health Classification
- Hepatitis C Virus (HCV) Dataset

A. Hepatocellular Carcinoma data set (HCC data set)

In recent years, the role of machine learning algorithms in supporting HCC medical work can not be ignored. In terms of prognosis, Santos et al. utilized a cluster-based oversampling method based on the K-means clustering SMOTE algorithm to build a model for predicting the 1-year survival of HCC patients, and the model achieved the best classification efficiency of 75.19%.[8] HCC Data set is made publicly available in the UCI Machine Learning repository A university hospital in Portugal provided the HCC data set, which contains information on 165 actual patients who had been diagnosed with the disease as well as a demographic, risk factor, laboratory, and overall survival characteristics. 49 features make up the data set. With 26 qualitative and 23 quantitative characteristics, this data set is varied. Overall, 10.22% of the total data set is missing data, and just eight patients (4.85%) have complete data across all fields. The survival rate at one year was the target variable, and it was encoded as a binary variable: 0 (dies) and 1. (lives). There is also some degree of imbalance in the class data. Data set contains different features like Gender,Symptoms,Alcohol,Age,Diabetes,Haemoglobin,Platelets and Albumin etc.,[8][9][10]

B. Fetal Health Classification

The descriptions and features of the CTG data sets pertaining to pregnancy problems are introduced in this subsection. The UCI Machine Learning Repository's CTG databases were utilized to compile the CTG data sets that were used in this study [11]. These databases provide information on FHR readings and uterine contraction characteristics throughout pregnancy based on Cardiotocograms, which was provided in September 2010 by the Faculty of Medicine at the University of Porto in Portugal and the Biomedical Engineering Institute, Porto, Portugal. These data sets, which have a steadily growing dataset size, were gathered based on clinical pregnancy cases in 1980 and on an irregular basis between 1995 and 1998. The CTG dataset contains 2,126 clinical occurrences indicating various pregnancy problems on fetal cardiotocograms. The fetal cardiotocogram clinical cases were automatically processed, and their corresponding diagnostic characteristics were measured. Three experienced obstetricians provided consensus categorization labels to each of these clinical occurrences and categorized them according to a morphologic pattern. In the CTG dataset, each clinical instance has one fetal state, 21 input characteristics, and one multiclass attribute. The 10 target classes are included in the multiclass morphologic patterns represented by the multiclass property. This multiclass property is also represented by an integer with a value between ”1” and ”10,” where each integer corresponds to a different morphological pattern during pregnancy.[15] This multiclass property is also represented by an integer with a value between ”1” and ”10,” where each integer corresponds to a different morphological pattern during pregnancy. Each of the three classes—normal, suspicious, or pathologic cases—is given a classification for the fetal condition. By removing the questionable class category in the fetal state, the CTG dataset may be utilized to create classification and prediction models based on the 10-class, 3-class, or even 2-class classification studies. In earlier works [16]–[19], a number of machine learning–based classification models were constructed based solely on a binary categorization of the fetal state in terms of normal and pathologic instances, excluding all questionable cases in fetal state.[15][20] In this paper we did a 3-class classification study.

C. Hepatitis C Virus (HCV) Data set

Hepatitis C Virus (HCV) is a worldwide illness that affects the human population. It's a blood-borne infection that spreads by direct contact with contaminated blood or blood-containing bodily fluids. Hepatitis C is a worldwide illness, according to the World Health Organization (WHO).

A team from the Institute of Clinical Chemistry at the Medical University of Hannover (MHH), Hannover, Germany, collected the HCV data set from a published study on hepatitis C patients. It offers information on 615 genuine HCV patients, which would include demographics, risk factors, test results, and overall survival. The data set has 14 attributes, which include Patient ID, Category, Age, Sex, and 10 laboratory data attributes.

The target attribute for classification is Category (blood donors(0) vs. Hepatitis C (including its progress ('just' Hepatitis C(1), Fibrosis(2), Cirrhosis(3).[12]

III. Data Analysis and Preprocessing

A. HCC data set

The data set lacked some information. Different problems arise when data are missing. Data absence reduces the power of statistics. The assessment of metrics may incline as a result of lost data. It may reduce the data samples' degree of representatives [13]. The missing data in the data set had the form of "?" First, we substituted "NaN" (Not-a-Number) values for the "?" values.[9][13]

Imputation substitutes a number for the absent esteem.The accepted approach is imputation, and it typically works flawlessly. However, attributed esteems may be significantly higher or lower than their genuine esteems. In this work, we used the mean value imputation technique for non categorical data and mode value imputation technique for categorical data.The process of bringing all the features to the same scale is called feature scaling. All of the elements in this work have been standardised. The outcome of standardisation is that the features are re scaled to ensure that the mean and standard deviation are, respectively, 0 and 1. The standardisation application principle is to divide the difference between the sample(x) and mean(x) by the standard deviation (x).Standardisation is performed on the data.

When models are conveyed over realised classes in an imbalanced or biased manner, it is referred to as an imbalanced classification issue. When there is one model in the minority class and hundreds, thousands, or millions of models in the majority class or classes, the distribution can change from having a moderate tendency to having an excessive stiffness. As the majority of the machine learning methods used for classification were designed around the suspicion of an equal number of models for each class, imbalanced classification serves as a test for prediction. This result in models that openly fail to accurately predict the future for the minority class.

The oversampling calculation known as SMOTE, or "Synthetic Minority Over-sampling Technique," uses the concept of nearest neighbours to create its enhanced data. Synthetic data is intentionally generated false information that closely resembles the form or projections of the data it is intended to enhance. A synthetic data generator produces data that is similar to the existing model rather than just replicating the data to create new models. SMOTE shines when it creates synthetic data. The computation gathers the K-nearest Neighbours of each observation that belongs to the minority class and creates another instance of the minority label at a random location along the line between the present observation and its closest neighbour.The data set's class variable exhibited an unbalanced distribution of data. As a result, we have implemented SMOTE to address this issue.[9][14]

B. Fetal Health Classification

There are no null values in the data. The data has been analyzed using a number of graphs, including the Count plot, Correlation Plot, Implot, Swarm and Boxen Plot. We learned from the count plot that the data is unbalanced whicg can be seen in "Fig.1", and performance measurements including the confusion matrix, precision, recall, and F1 Score have been taken into consideration to provide more useful insights. According to the correlation matrix, "accelerations","prolongued decelerations", "abnormal short term variability", "percentage of time with abnormal long term variability" and "mean value of long term variability" are the parameters that have a greater association with fetal health.

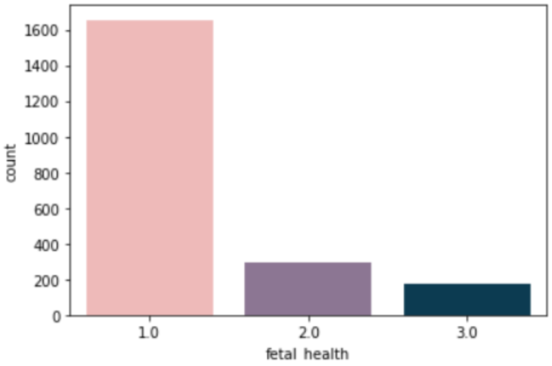

Fig. 1. count plot of the data.

Swarmplot was used to teach us about outliers; one example is shown in "Fig.2". However, removing them results in overfitting, and because this is the outcome of a CTG report in our case, it is unlikely that the data input was erroneous. If the outlier does not belong to the population being studied, it can be correctly eliminated. In this case, the foetus is the major concern, and experts have given it a classification. Since we followed the expert opinion, we didn't delete it if these were a natural part of the population under consideration.

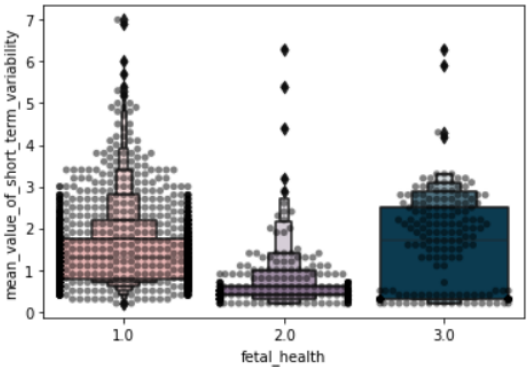

Fig. 2. depicting outliers in the data.

C. Hepatitis C Virus (HCV) Data set

When we first looked at the data set information, we came to the conclusion that there were a few nameless attributes in the data that were unneeded, so we eliminated them. Then, among the Data Attributes themselves, we had a number of Data Types, including integers, objects, and floats. Almost all machine learning algorithms require numerical data for learning because the dataset contains categorical variables, as was previously noted. This implies that we should immediately encode categorical data into a numeric representation. One of the widely used methods for converting labels into numeric values so that machines can understand them is label encoding. For the categorical variables, label encoding has been done. To convert the labels into a numeric format, label encoding was performed. The Numeric data type has replaced the Object data type.

The data was then shown using a variety of visualisation approaches, such as correlation matrices, scatter plots, box plots, and other visualisations, before being over sampled to standardise the data as a whole.

IV. System architecture

A. Block Diagram

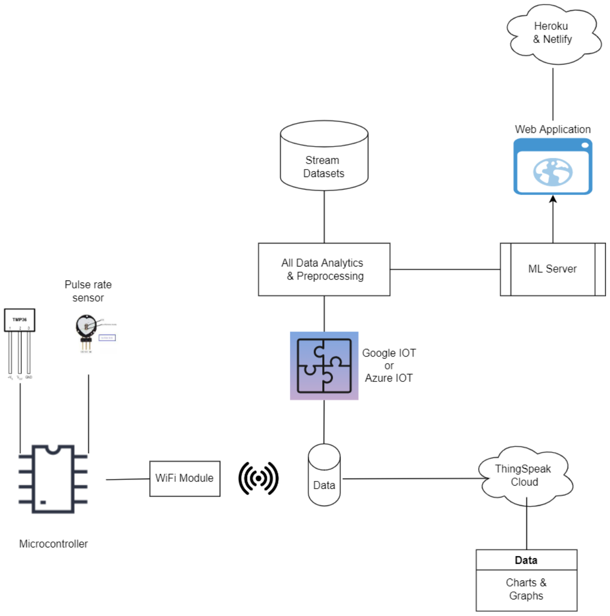

Fig. 3. Proposed System architecture

Electrocardiogram (ECG), pulse rate, pressure, temperature, and position detection are the 5 parameters we are using. The system has two circuits: the transmitting circuit is with the patient, and the doctor can see the receiving circuit.

These sensors on the Arduino board are now sending data to the microcontroller, which then sends it to Google/Azure IoT via a WiFi module. The same Cloud is used to store the data. We use ThingSpeak IoT to generate charts and graphs out of that data.

Now that we have added Stream Datasets in addition to this IoT data, both of them are preprocessed. Now, we send this data to the ML server, where it is stored and put to the test.

In order to visualize the created model and provide a platform for the development of an interaction between the user, the model, and the medical team, a user-friendly and user-responsive web application will be created. The same will be deployed in Heroku and Netlify Cloud.

B. Incremental learning architecture

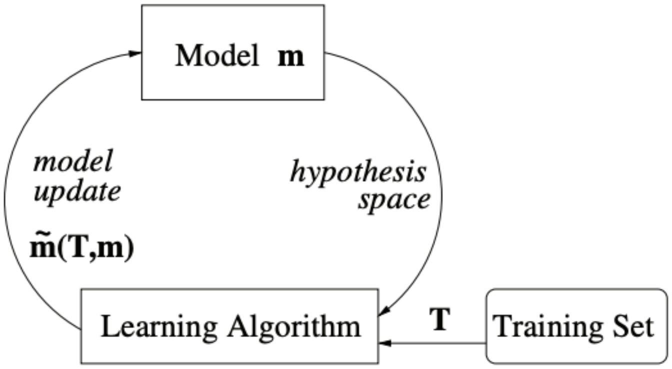

Fig. 4. Incremental learning architecture

The algorithms shown in the below figure include incremental learning algorithms, which add a new sample at each iteration; relevance feedback algorithms, which add new positive and negative examples at each iteration; and algorithms where training is accomplished through a series of iterations using a fixed training set. The learner replaces the current model m with a fresh model, m(t,m), chosen from the training set t, during each iteration. The hypothesis space of the learner given the current model is the set of all m(t,m)'s the learner has chosen as t changes over T, if T represents the collection of all potential training sets.

V. ML Algorithmic Analysis

A. Batch Learning Algorithms

1) Logistic Regression(LR): In the supervised learning technique, the logistic regression is one of the most commonly used ML algorithms [22]. It is a forecasting technique that makes use of a group of independent variables to anticipate the value of a categorical dependent variable.The output of a categorical dependent variable is forecasted using logistic regression. As a result, the output must be either categorical or discrete. It may be Yes or No, 0 or 1, true or false, and so on, but probabilistic values between 0 and 1 are given instead of precise values like 0 and 1[21]Here in this research we have implemented Logistic Regression on HCC Data, Fetal Health and HCV Data.

2) XGBoost(XGB): A gradient boosting technique is used in the decision tree-based group machine learning process known as XGBoost to solve classification and regression issues. This process begins with a simple model to create a forecast. The loss of this model is then calculated. Another brand-new model is taught to reduce this. For classification, this model is being included in the ensemble. Decision trees are created in serial formation for this calculation. Weights make up a large portion of XGBoost.All of the independent factors receive weight transfers before being passed to the decision tree, which predicts outcomes. The decision tree expands the weight of elements that it incorrectly predicted, and these factors are subsequently transmitted to the following decision tree. These different classifiers contribute to the ensemble to produce a classifier that is more reliable and accurate. [9] Here in this research we have implemented XGBoost on HCC Data, Fetal Health and HCV Data

3) Random Forest(RF): The random forest is a machine learning technique for guided learning [23]. It constructs a "forest" from a selection of trees that have been mostly prepared for the "bagging" technique. The bagging technique is fundamentally justified since mixing several learning models enhances the final result. The random forest creates a large number of different trees and then combines them to provide a more accurate and reliable representation. It has the advantage of addressing the arrangement and relapse issues that afflict the majority of existing ML frameworks. Another notable aspect of the random forest technique is the ease with which the general significance of each component in the estimate may be determined. The flexibility of random forest is one of its most alluring features. It may be utilized for both relapse detection and grouping tasks, and the overall weighting given to information characteristics is readily apparent. Additionally, it is a beneficial approach since the default hyper parameters it employs often give unambiguous expectations. Understanding the hyper parameters is critical, since there are relatively few of them to begin with. Overfitting is a well-known problem in ML, although it occurs seldom with the arbitrary random forest classifier. If there are sufficient trees in the forest, the classifier will not overfit the model[21]. Here in this research we have implemented Random Forest on Fetal Health,HCV Data

4) Decision Tree(DT): This classifier [24] seems to recursively divide the example space. It is a predictive paradigm that acts as a mapping between the characteristics of an item and their values [25]. It regularly splits each potential data result into parts. Each nonleaf node corresponds to a feature experiment, each branch to the outcome of the experiment, and each leaf node to a judgment or classification [26]. The root node of the tree, which is at the very top, reflects the most often used prediction model. The decision node and the leaf node are the two nodes in a decision tree. The choice nodes are used to make those selections and have numerous branches, whereas the leaf nodes are the result of those choices and contain no additional branches. The outcomes of the tests or judgments are contingent on the dataset's properties.The choice tree is easy to comprehend since it replicates the phases that a person goes through while making a real-world decision. It may be very beneficial in resolving issues with decision-making. Consider all potential solutions to an issue[28][21]. Here in this research we have implemented Decision Tree on Fetal Health Data and HCV Data.

5) Support Vector Classifier(SVC): The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations.

we've implemented the SVC on HCC Data, Fetal Health and HCV Data.

B. Incremental Learning Algorithms

1) Hoeffding Tree Classifier(HT): Hoeffding tree classifier [28] is a member of decision tree family developed to solve streaming classification problems. It is based on the assumption that the distribution that generates the examples does not change over time. In Hoeffding tree algorithm, Hoeffding bound is used in tree induction to determine how many examples are needed to achieve certain level of confidence for the tree split purpose. Instead of needing a large number of examples to assess the split evaluation function, Hoeffding bound guarantees that it is possible to generate the tree using only a limited number of examples. Hoeffding tree is different compared to the other classical decision tree learners like ID3, C4.5, and CART. Traditional decision trees assume that all training examples are stored in the main memory whereas, in Hoeffding tree, the tree update process is conducted with each arrived data before the data is discarded. An important property of Hoeffding tree algorithm is that it can generate tree that is asymptotically close to the one generated via batch learning [28]. There are several extensions of Hoeffding tree such as CVFDT [29], VFDTc [30], UFFT [31], and Adaptive Hoeffding tree [32][27]. In this research we have implemented the HT on HCC Data, Fetal Health, and HCV Data.

2) Incremental Naive Bayes(INB): Carries out traditional Bayesian prediction while erroneously assuming that all inputs are independent and fitting the data incrementally. A classifier algorithm notable for its simplicity and cheap processing cost is incremental Naive Bayes. The trained Naive Bayes classifier accurately predicts for each incremental occurrence the class to which it belongs given n distinct classes. Its benefit is that it can maintain the parameters contained in the initial training data.Here we have implemented Incremental Naive Bayes on HCC Data.[7].In this research we have implemented implemented the INB on HCC Data, Fetal Health, and HCV Data.

3) Adaptive Random Forest Classifier(ARF): Without additional hyper-parameter adjustment, Adaptive Random Forest can achieve good classification in data streams with various features. It will be helpful for both practical applications and as a benchmark for upcoming algorithm ideas in the field because it is a sustained off-the-shelf learner for the difficult problem of evolving data stream classification.Re-sampling, randomly picking feature subsets for node split and drift detectors per base tree, which prompt selected resets in response to drifts, are the three most crucial components of Adaptive Random Forest [7]. Additionally, it enables the training of background trees, which begin training when a warning is identified and then the warning develops into a drift. In this research we have implemented ARF on HCC Data and Fetal Health

4) KNN ADWIN Classifier: K-Nearest Neighbors classifier and the ADWIN change detector Due to its resistance to concept drift, this Classifier outperforms the standard KNN Classifier. It controls the sample window size by selecting which samples to keep and which to discard using the ADWIN change detector. The main differences between this class and the standard KNN Classifier are that it maintains a variable size window rather than a fixed size one and that it updates the Adwin algorithm with each partial fit call. The window of fixed size approach is the simplest principle, and the user often chooses the window size. A window of fixed size method is an excellent option if you are aware of the rate of change.However, the user is frequently forced to make a trade-off between selecting a tiny window size (which quickly adapts to the current distribution) or a big window size[6]. According to the drift's length, the adaptive window technique modifies the window's size.In this research we have implemented KNN ADWIN on HCC Data and Fetal Health and HCV Data

5) Learn++.NSE Classiffier(LPP.NSE): Learn++. NSE is an algorithm from the Learn++ family. Learn++. For incremental learning from non-stationary environments (NSEs) where the underlying data distributions fluctuate over time, NSE is an ensemble of classifiers. It picks up new information from successive batches of data that drift at a constant or variable pace, add or remove concept classes, and cycle through drift[7].Learn++. NSE is an ensemble-based batch learning technique that use weighted majority voting. The weights are constantly changed in light of the classifiers' time-adjusted mistakes on both the present environment and the environment from the past. This method enables the algorithm to detect changes in the underlying data distributions and respond appropriately, as well as to the potential reappearance of an earlier distribution. It uses just recent data for training and a passive drift detection technique. Numerous non stationary situations, such as abrupt idea changes, quick or slow, gradual or abrupt, cyclical, or even variable rate drift, can be handled by it. Additionally, it is one of the few algorithms that can manage the deletion of an existing class or the insertion of a new class[5].In this research we have implemented Learn++.NSE on HCC Data

VI. Results and Discussions

Accuracy, Precision, Recall, Kappa, and F1 Score were the evaluation measures we used to assess the effectiveness of our deployed system.Kappa is a statistical measure that assesses the consistency of two raters' ratings of the same quantity and shows how often the raters agree.Kappa can have a value that is lower than 0. (negative). A score of 0 indicates that the raters' opinions differ randomly, whereas a score of 1 indicates that they are entirely in accord. Consequently, a score smaller than 0 indicates that there is less agreement than could be predicted by chance.

A. HCC Data set

This Data generated a good result for XG Boost with an accuracy rate 78.12%,F1 Score 84.4%,Precision 82.6% and Recall of 86.3%. The other batch learning algorithm Logistic Regression produced 65.62% Accuracy, 71.7% F1 Score,60.8% Precision and 87.5% Recall.Remaining Batch Learning Algorithms like Decision Tree , Random Forest and Support Vector Classifier are mentioned in the summary table. While Batch learning Algorithms like XGBoost Performed well on the data.Incremental Learning Algorithms also have generated a credible score with good impact on data where KNN Adwin (Adaptive Window Size) surpassed all other incremental algorithms with 62% Accuracy, 0.16 Kappa Score,71.8% F1 Score,67.2% Precision and 77% Recall. Incremental Naive Bayes also performed well with 60.6% Accuracy, 0.20 Kappa Score,65.5% F1 Score,71.6% Precision and 65.5% Recall. The Kappa Scores of remaining Incremental Algorithms like Adaptive Random Forest,Hoeffding Tree Classifier generated negative scores whereas Learn++.NSE produced an exact 0 kappa score on the HCC Data and even produced a less accuracy score of 38.06% when compared to all other algorithms.The Accuracies of the implemented algorithms on the respective data sets are summarised in table 1.

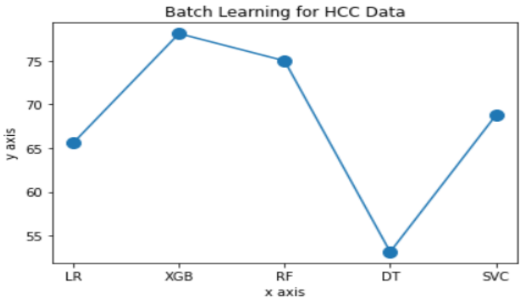

Fig. 5. Comparison of Batch learning algorithms

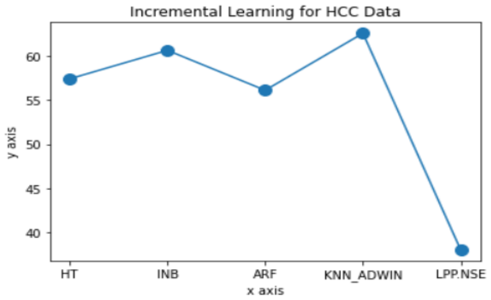

Fig. 6. Comparison of Incremental learning algorithms

B. Fetal Health Classification

Random Forest and XGBoost both performed exceptionally well on the data set, with an accuracy rate of 94.51% and 94.04%, respectively. Decision Tree came in second with an accuracy rate of 93.88%, which is completely understandable given that Random Forest's feature space is divided into more and smaller regions and its trees are more diverse than Decision Tree's. For the best did Random Forest the precision is 94.40%, Recall is 94.51% and F1 Score is 94.51%. The ROC curve of the Randomforest is in fig.7. And also the comaparison of Batch Algorithms are in fig.8 and Incremental algorithms are in fig.9.

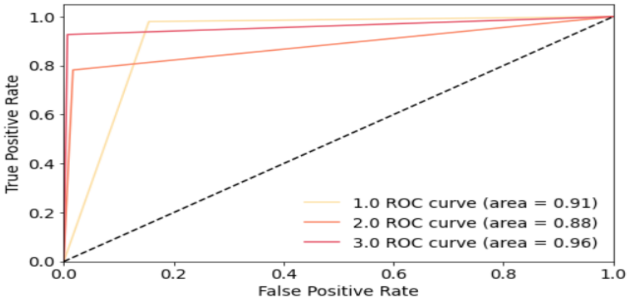

Fig. 7. ROC Curve for Random Forest Classifier.

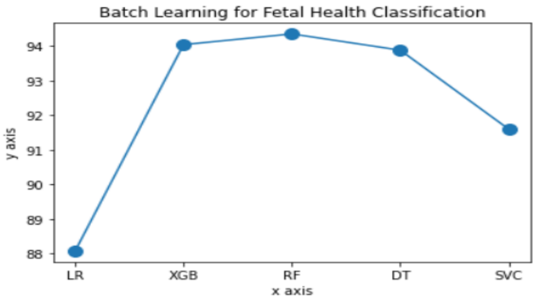

Fig. 8. Comparison of Batch learning algorithms

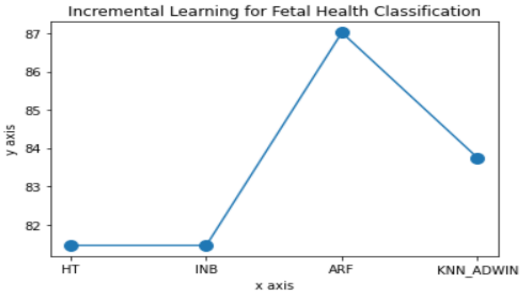

Fig. 9. Comparison of Incremental learning algorithms

With an accuracy of 87.02%, Adaptive Random Forest surpassed all other implemented models in the scenario of Incremental Learning, followed by KNN ADWIN with 83.75%. Batch learning or offline learning performed well overall, however incremental learning also performed well on this dataset.The Accuracies of the implemented algorithms on the respective data sets are summarised in table 1.

C. HCV Dataset

With accuracy rates of 99.6 %, batch algorithms like XGBoost and Random Forest performed brilliantly on this data set. With an accuracy of 86.99 %, Logistic Regression performed wonderfully.Speaking specifically of LR, the precision was found to be 83.88 %, recall to be 86.99 %, and F1 to be 85.24 %. SVC accuracy measured at 90.90% and Decision Tree accuracy measured at 99.3%.

With an accuracy of only 12 percent, KNN Adwin appears to be performing poorly on this dataset.Other algorithms which were also used are Hoeffding Tree Classifier with an accuracy of 3.755 %, Naive Bayes with an accuracy of 5.79 %, Adaptive random Forest with an accuracy of 5.38 % which in terms of incremental learning might be meritless at this point.

This particular dataset is more well suited for batch learning algorithms over Incremental learning algorithms. This could be as a result of challenges with training on a series of limited data from new tasks, which results in serious overfitting problem. Below table represents the accuracy scores summarised of all implemented algorithms with respect to data sets. Based on the data in the datasets we applied only certain algorithms to the respective dataset. In the summarised table '-' represents the algorithm not applied to the particular dataset.

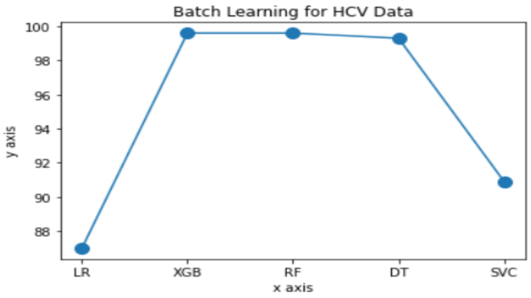

Fig. 10. Comparison of Batch learning algorithms

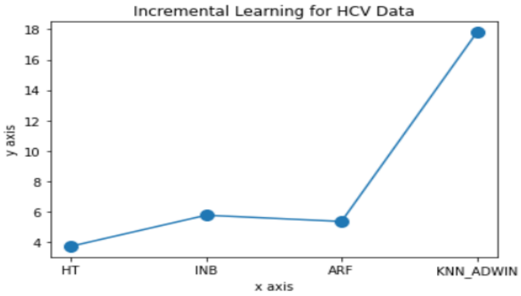

Fig. 11. Comparison of Incremental learning algorithms

Conclusion

The reason to use Incremental learning over Traditional Machine learning is that, the goal of incremental learning is to create artificially intelligent systems that can continuously learn to solve new problems from fresh input while retaining the knowledge acquired from earlier solved problems.

We acquired three datasets: one each for HCC, fetal health, and HCV. Preprocessing of the data has been completed. According to the need, well-known batch learning algorithms like

TABLE I
TEST ACCURACIES(%) OF RESPECTIVE ALGORITHMS OF DATA SETS

Algorithms	***HCC***	***Fetal***	***HCV***
LR	65.62	88.08	86.99
XGB	78.12	94.04	99.6
RF	75.0	94.35	99.6
DT	53.12	93.88	99.3
SVC	68.75	91.6	90.9
HT	57.42	81.46	3.755
INB	60.65	81.46	5.795
ARF	56.13	87.02	5.387
KNN ADWIN	62.58	83.75	17.795
LPP.NSE	38.06	-	-

Logistic Regression, XGBoost, RandomForest, and SVC have been used. Likewise used are incremental learning algorithms like KNN ADWIN, Hoeffding Tree, Adaptive Random Forest, and Naive Bayes. Algorithms of the same learning type are compared to one another.

In our future work, we plan to design a incremental model to predict the patient health status from the real time data. We also plan to propose a strategy to resolve Concept Drift.

ACKNOWLEDGMENT

We are incredibly grateful to Dr. Simi Surendran for her feedback and patience, which are both crucial. She kindly shared her knowledge and expertise, without which we would not have been able to embark on this journey.

REFERENCES

[1] Chen, Hongge, and Duane Boning. "Online and incremental machine learning approaches for IC yield improvement." 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE, 2017.

[2] Raghuraman, Kirthanaa, et al. "Online incremental learning algorithm for anomaly detection and prediction in health care." 2014 International Conference on Recent Trends in Information Technology. IEEE, 2014.

[3] He, Haibo, et al. "Incremental learning from stream data." IEEE Transactions on Neural Networks 22.12 (2011): 1901-1914.

[4] Lu, Jie, et al. "Learning under concept drift: A review." IEEE Transactions on Knowledge and Data Engineering 31.12 (2018): 2346-2363.

[5] Elwell, Ryan, and Robi Polikar. "Incremental learning of concept drift in nonstationary environments." IEEE Transactions on Neural Networks 22.10 (2011): 1517-1531.

[6] Iwashita, Adriana Sayuri, and Joao Paulo Papa. "An overview on concept drift learning." IEEE access 7 (2018): 1532-1547.

[7] https://scikit-multiflow.readthedocs.io/en/stable/api/api.html

[8] Dhillon, Arwinder, and Ashima Singh. "Machine learning in healthcare data analysis: a survey." Journal of Biology and Today's World 8.6 (2019): 1-10.

[9] Akter, Laboni, and Md Milon Islam. "Hepatocellular carcinoma patient's survival prediction using oversampling and machine learning techniques." 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2021.

[10] Hepatocellular Carcinoma (HCC) Dataset http://archive.ics.uci.edu/ml/datasets/HCC+Survival

[11] Fetal Health Classification https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification

[12] Hepatitis C Virus (HCV) Dataset https://archive.ics.uci.edu/ml/datasets/HCV+data

[13] Kang, Hyun. "The prevention and handling of the missing data." Korean journal of anesthesiology 64.5 (2013): 402-406.

[14] Rahman, M. Mostafizur, and Darryl N. Davis. "Addressing the class imbalance problem in medical datasets." International Journal of Machine Learning and Computing 3.2 (2013): 224.

[15] Miao, Julia H., and Kathleen H. Miao. "Cardiotocographic diagnosis of fetal health based on multiclass morphologic pattern predictions using deep learning classification." International Journal of Advanced Computer Science and Applications 9.5 (2018).

[16] P. A. Warrick, E. F. Hamilton, R. E. Kearney, and D. Precup, "A machine learning approach to the detection of fetal hypoxia during labor and delivery," Proceedings of the Twenty-Second Innovative Applications of Artificial Intelligence Conference, pp. 1865-1870, 2010.

[17] Z. Comert and A. F. Kocamaz, "Comparison of machine learning techniques for fetal heart rate classification," Special issue of the 3rd International Conference on Computational and Experimental Science and Engineering, Vol. 132, pp. 451-454, 2017.

[18] C. Sundar, M. Chitradevi, and G. Geetharamani, "Classification of cardiotocogram data using neural network based machine learning technique," International Journal of Computer Applications, Vol. 47, No. 14, pp. 19-25, June 2012.

[19] M. Arif, "Classification of cardiotocograms using Random Forest classifier and selection of important features from cardiotocogram

[20] C. Gribbin and J. Thornton, Critical evaluation of fetal assessment methods, In: James DK, Steer PJ, Weiner CP editor(s), High Risk Pregnancy Management Options, Elsevier, 2006.

[21] Alam, M.T., Khan, M.A.I., Dola, N.N., Tazin, T., Khan, M.M., Albraikan, A.A. and Almalki, F.A., 2022. Comparative Analysis of Different Efficient Machine Learning Methods for Fetal Health Classification. Applied Bionics and Biomechanics, 2022.

[22] "Logistic Regression in Machine Learning," https://www.javatpoint.com/logistic-regression-in-machine-learning.

[23] Donges, Niklas. "A complete guide to the random forest algorithm." Built In 16 (2019).

[24] M. A. Mohammed, M. K. A. Ghani, R. I. Hamed, and D. A. Ibrahim, "Analysis of an electronic methods for nasopharyngeal carcinoma: prevalence, diagnosis, challenges and technologies," Journal of Computational Science, vol. 21, pp. 241–254, 2017.

[25] F. Moreno-Seco, L. Micó, and J. Oncina, "A modification of the LAESA algorithm for approximated k-NN classification," Pattern Recognition Letter, vol. 24, no. 1-3, pp. 47–53, 2003.

[26] M. A. Mohammed, M. K. A. Ghani, R. I. Hamed, and D. A. Ibrahim, "Review on nasopharyngeal carcinoma: concepts, methods of analysis, segmentation, classification, prediction, and impact: a review of the research literature," Journal of Computational Science, vol. 21, pp. 283–298, 2017.

[27] Pham, X.C., Dang, M.T., Dinh, S.V., Hoang, S., Nguyen, T.T. and Liew, A.W.C., 2017. Learning from data stream based on random projection and hoeffding tree classifier. In 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (pp. 1-8). IEEE.

[28] P. Domingos and G. Hulten, "Mining high-speed data streams," in Proceedings of the KDD conference, 2000, pp. 71–80.

[29] Geoff Hulten, L. Spencer and P. Domingos, "Mining time-changing data streams," in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001, pp. 97–106.

[30] J. Gama, R. Fernandes and R. Rocha, "Decision trees for mining data streams," Intelligent Data Analysis, vol. 10, no. 1, pp. 23–45, 2006.

[31] J.Gama, P. Medas and P. Rodrigues, "Learning decision trees from dynamic data streams," Journal of Universal Computer Science, vol. 1, no. 8, pp. 1353–1366, 2005.

[32] R. G. A. Bifet, "Adaptive parameter-free learning from evolving data streams," in Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII, 2009, pp. 249–260.

Signature: