Attention-Enhanced BiLSTM for Score Forecasting in Parkinson's Disease

A thesis submitted towards the partial fulfilment of the requirements for the degree of Master of Engineering in Biomedical Engineering

Course affiliated to Faculty of Engineering and Technology, Jadavpur University

Submitted By-

Sayantani Chakraborty
ROLL NO.:002030201009
REGISTRATION NO.: 154633 of 2020-2021
Examination Roll No. M4BMD22008

Under the guidance of:-
Dr. Anasua Sarkar

&

Co-guidance of:-
Dr. Piyali Basak

School of Bioscience and Engineering
M.E. in Biomedical Engineering course affiliated to Faculty of Engineering and Technology
Jadavpur University
Kolkata-700032
India

CERTIFICATE OF RECOMMENDATION

We hereby recommend that the thesis entitled "Attention-Enhanced BiLSTM for Score Forecasting in Parkinson's Disease" carried out under my supervision by Sayantani Chakraborty may be accepted in partial fulfilment of the requirement for awarding the Degree of Master in Biomedical Engineering of Jadavpur University. The project, in our opinion, is worthy for its acceptance.

Dr. Anasua Sarkar
(Thesis Advisor)
Assistant Professor
Computer Science and Engineering
Jadavpur University
Kolkata-700032

Director
School of Bioscience and Engineering
Jadavpur University
Kolkata-700032

Dean
Faculty Council of Interdisciplinary Studies, Law and Management
Jadavpur University
Kolkata – 700032

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains literature survey and original research work by the undersigned candidate, as part of her Master of Engineering in Biomedical Engineering studies during academic session 2020-2022.
All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.
I also declare that, as required by this rules and conduct, I have fully cited and referred all material and results that are not original to this work.

NAME: Sayantani Chakraborty
EXAM ROLL NO.: M4BMD22008
CLASS ROLL NO.: 002030201009
THESIS TITLE: Attention-Enhanced BiLSTM for Score Forecasting in Parkinson's Disease

---

Sayantani Chakraborty
M.E.,School of Bioscience & Engg
Jadavpur University
Kolkata-700032

CERTIFICATE OF APPROVAL

The forgoing thesis is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it is submitted.

Dr. Anasua Sarkar
(Thesis Advisor)
Assistant Professor
Computer Science and Engineering
Jadavpur University
Kolkata-700032

External Examiner

ACKNOWLEDGEMENT

I owe a deep sense of gratitude to my respected thesis advisors Dr. Anasua Sarkar Asst. Professor, Computer Science & Engineering Department, Jadavpur University for her esteemed guidance, invaluable suggestions, constant encouragement and affection at every stage of the entire tenure of the project without which I could not have finished the work. It has been my proud privilege to work under her guidance.

I would also like to express deep felt gratefulness to my Co-guide Dr. Piyali Basak Associate Professor of School of Bioscience and Engineering for her kind support and guidance.

I would also like to thank all the nonteaching staffs, all my batch mates, my seniors, and all my junior friends for their kind co-operation.

Last, but not the least, I wish to express my profound gratitude and my deep feelings for my family who have been the constant source of my energy, inspiration and determination for going ahead with my academic pursuit.

Sayantani Chakraborty
School of Bioscience & Engg.
Jadavpur University
Kolkata-700032

Contents

1 Abstract

Parkinson' Disease is the most progressive neurological disease that is caused because of the dying of the neurons or brain cells.It is the most common neurological disease caused around the globe, with most patients in the age greater than 60[1].The disease is characterised with tremors,instability in the walking or posture even problem in sleeping along with speech problems. Speech is one of the earlier trait in diagnosing the disease at the earliest stage[9]. So in this study the speech characteristics or features are taken advantage of to forecast the disease at the earliest.Using a patient dataset that has a clinical PD grading based on speech characteristics could slow the progression of PD by offering a computational prognostic tool for the condition.Based on previously and now recorded speech, it can assist a person with PD in tracking the development of unique symptoms they are currently experiencing. This study uses recurrent neural network, Bidirectional Long-Short term memory or BiLSTM to forecast the time series based output. The traditional BiLSTM framework is used as base model and an attention layer is used on top of that to increase the efficiency or performance of forecasting.The model proposed is called attention-BiLSTM.

The performance of the model is compared with other tradition recurrent neural models like LSTM and BiLSTM.Regression models are commonly used in the forecasting models. So, the performance of model is also compared with that of multiple linear regression.A ridge regression is an advanced form of traditional linear regression is also run on the the dataset and is also used for the comparison of the attention based BiLSTM model.Different performance metrics are used to study the performance. Though RMSE is the metrics that is widely used in literature for the performance analysis in the forecasting but in this study along with the RMSE, other metrics like MSE and MAE are also used in the study.The proposed attention enhanced BiLSTM model is showing a result of 7.58%,4.36% and 2.08% for MAE,MSE and RMSE respectively, that is better than the comparing/base models in the study.

2 Introduction

Machine learning's crucial field of time series forecasting is frequently ignored. The fact that so many prediction issues have a temporal component makes it crucial. Due to the time component making time series problems more challenging to manage, some issues go unresolved. An observational dataset is what a typical machine learning dataset consists of. In typical machine learning datasets, time does matter.

2.1 Time Series Forecasting

When new data is present, predictions are made even though the final result may not be known for some time. Predictions are made about the future, but virtually always treat all past observations equally. Perhaps employing relatively minimal temporal dynamics, such as using only the most recent year of observations rather than all available data, we can avoid the problem of "concept drift. "Different is a time series dataset. Time series introduce a time dimension, which is an explicit order dependence between observations. This additional dimension serves as a limitation as well as a framework for more data. Depending on whether we want to analyze a dataset or make predictions, our objectives are different. Although it is not necessary, understanding a dataset known as time series analysis can result in a significant technical investment of time and knowledge that is not immediately related to the desired purpose, which is future forecasting. A time series is modelled in descriptive modelling, also known as time series analysis, to identify its components, perhaps with additional information.

The analysis of time series is the main issue when utilizing classical statistics. In order to truly understand the underlying causes of an observed time series, time series analysis involves creating models that best capture or define the time series. In this area of research, the "why" of a time series dataset is sought after.

Making assumptions about the dartaset structure and breaking down the time series into its constituent parts are frequent steps in this process. The accuracy of a descriptive model's description of all available data and the interpretation it offers to better understand the issue domain determine the model's quality. Time series analysis' main goal is to create mathematical models that can generate believable descriptions from sample data.

Extrapolation is the term used in conventional statistical processing of

time series data to describe making predictions about the future. Forecasting is to predict future value from the historical data using models.

Time series models are typically divided into two broad categories: Traditional time series models and machine learning-ML models. Traditional models are further divided into univariate models (ARIMA, SARIMAX) and multivariate models (Vector AutoRegression). ML models includes neural network regressors or regression models. ML models can make direct predictions based on the horizon. It is tough to extend. The training data increases linearly as we have more horizons to predict. ML models are easy to get right. It can also add time varying variables as features.

A time series is a observation set which are taken as equal intervals of time. This is used for future value prediction which are based on the previous observed value. Over here at X-axis have time and, on Y-axis, the data magnitude. i.e., if we try to plot time series plot, on the x co-ordinate we always get the time which is divided in equal intervals. We cannot create a time series where we have one data point at week level and other are different scale. This should be in equal interval say all data points in days or weeks or years and so on. So that is the constant thing that a time series require.

There is much importance of time series analysis. In business forecasting, the past data points define what is going to happen in the future. For example, a company is trying to predict what will be the price of it's share in the stock market the next day. However, time series analysis is not only limited to finance or retailing but is applicable almost everywhere. Time series also help us to analyze a past behavior. Here we can analyze say in which month the sales went up or when was the dip, so here one can easily understand or retrospect past data or it's behavior. This can be understood with respect to time. For example, some festival is there, and a company is selling sweets, so the sales will increase during the festival, then one need to think about the seasonality part. Time series analysis also helps to plan future operations as well. Therefore, one can analyze the past and then can forecast the future using this algorithm that is time series analysis. Apart from all these we can also evaluate current accomplishments. This means one can determine which goals have been met in the current scenario. Let's say one has predicted or planned to sell a certain amount of goods in a day but didn't do that. So, this also can be analyzed.

When not to apply time series analysis:

We cannot use time series analysis when the values are constant i.e., say the temperature of a particular place is 29 °C on Monday and the temperature

of the same place is 29°C on Tuesday and so on. Now, if we want to predict the temperature on the next coming days. In such cases where the values are constant as in this case the temperature is constant, time series cannot be applied. Similarly, if the values are in the form of functions, say we have sine of X or cos of X. In this case we have X value and one can get the value by just putting it in the function so there is no point of applying time-series analysis where one can calculate the values by just using the functions. We can however apply the analysis in these situations as well but again there is no point of applying it if we can easily do that.

2.2 Parkinson's Disease (PD)

PD is most general neuro-degenerative disorder, affecting people usually aged above 60 years. In India, there are around 10 lakhs instances each year. The risk is however incresed by age and the average age of diagnosis is sixty. In fact, people tend to think that PD is a disease of old age but some get the PD at 40 or even younger. It can last for years or for a lifetime. It is resulted by death of nerve cells in basal ganglia region of the brain. It is a progressive disorder. The Parkinson's disease is characterised by the movement or gait disorder. This often includes the tremors, stiffness, and slowness[18]. These can have dare consequences to the patient resulting in falls, reduced movement as well as dependence. Medication can help in relieving these symptoms and improves life quality but the correct dosage of the same calls for a system which determines the diseased state based on the wearables signals[2].

Most diseases can be find with test in labs. We need tests like that in PD detection, but they don't exists yet. Medical professionals look for medical history as well as real time physical examination of the subject.Their prime factors are motor disability but with that they also look for the features like sleep disorder, smelling loss or depression. In Parkinson's,cells that make dopamine stop producing, Dopamine is a signalling chemical that coordinates movements.This results in Parkinson's symptoms emerge.

Researchers believe that in most people PD is caused by genetic and environmental factors combination. But certain environmental factors like pesticide usage and injury are factors with an increased risk. Similarly, while genetic mutations are also a factor for the risk. However, the field of genetic is moving fast, tremendous research is genetics focused.The currently available PD medications can not slow or stop disease progression, but they

ease symptoms and helped the patient to continue doing much of what the patient have always done. The same goes for the surgical procedures. .

Along with movement change, Parkinson disease can affect speech. The voice of the patient may get hoarse or softer, pitch changes and the speech may become slurred. It is very difficult for the patient to accept these changes and seek help as often they are unaware that they are speaking more quietly than before. A speech difficulty can lead to reduction in self confidence and isolation. Social and family roles can also suffer as a result. Along with the consultation with a medical specialist or an appointment with a speech therapist, vocal exercises for speech difficulties such as regulating the voice volume and word articulation. In this work we have used the speech signal features along with motor scores for forecasting the future event of Parkinson episode.

Ninety percent of patients with early PD exhibit speech problems. As a result, this can be a powerful indicator for creating a trustworthy PD diagnosis[16]. Patients who have been diagnosed typically ought to be available in the clinic and attend routine check-ups[12]. Evaluation of PD symptoms primarily relies on human competence. Additionally, the UPDRS- Unified Parkinson's Disease Rating Scale, which necessitates motor skills evaluations assisted by qualified medical staff, is frequently used in traditional tracking of PD symptoms progression. Since PD patients frequently exhibit certain speech characteristics, audio recording data is useful for diagnosis. Smartphones can function as a mobile recording device, making it easy to remotely check on patients' health.

To address these issues, we provide a monitoring methodology that can predict the PD progression of 16 speech components and UPDRS scores over the course of a week or so. We have done a multivariate time series forecasting using various regression as well as LSTM and Bi-LSTM models for the forecasting using UPDRS and 16 speech signals features. Additionally, it is advantageous for patients' health be tracked remotely by a physician so they can avoid visiting the doctor for regular check-ups.

2.3 Recurrent Neural Network

The latest happening thing in sequences or analyzing sequences and forecasting is recurrent neural networks specially LSTMs, the most famous form of recurrent neural network. A traditional deep learning algorithm, say in image processing, use convolutional filters. These are two dimensional filters that are convolutional filters extracts various features from the input image in a very cascaded way in a step by step and a patch is applied. Features are extracted and then operations like max pools are done. Eventually before it goes into a fully connected layer, a whole bunch of features are generated as inputs to the dense layer or fully connected layer or feed forward layer and then the information goes through the forward direction and eventually what comes out is a probability, which again depends on the activation functions used. So all the information one need is in the input image and all the information is transferred one after the other and it goes in from left all the way to the right[19].

But in time series just the feed forward does not work and we need to send some information back. So the neuron actually has an idea of what happened in the past, in the history when information went through. So recurrent neural networks are different from regular neural networks, in the fact that they remember the past i.e, all RNNS remembers its past[17].

A question arises that why we even need recurrent neural networks when we have traditional neural networks. For example, we have a scenario – what's for dinner? And if we have a pattern which starts on Sunday where we have Japanese cuisine, then Thai, then Indian, then Chinese and so on and then again, the pattern repeats. So here say there is no correlation between days and cuisines. However, there is correlation between the Cuisine itself, between different data points. We say that auto correlation. This is why we need RNNS because regular rural networks we have X and Y and we model this X to predict Y. But here we have two models Y to predict Y i.e, here the decision of what to eat today depends on the cuisine of the day before and not depends on the weekday. So this sequential information is used by RNNS.

An RNN when unrolled, which is what happens during the execution i.e, during the training and prediction part. The input is coming (Xt) and there is a output (Ht) but then the information is travelling back within this cell, this is a RNN cell. If we unrolled this, it will be the same network, there are multiple copies of the same network and each of this will pass on information

to the next. It is like a list like structure or sequence like structure as in figure1.

There are many types of RNN: one to one, one to many, many to one and many to many. One to One is just like a regular neural network, like One input and One output, like image classification. One to Many is like image captioning, you supply an image, a cat in front of a car, so it is out putting a sentence a cat in front of a car explaining a scene. Many to One is exactly opposite. Which means there are multiple inputs and only one output, for example sentiment analysis. A bunch of tweets on a specific hashtag, it labels the tweets. Many to Many have many variations. This is for language translation, if one sentence is in English as input, it gives an output in German for example.

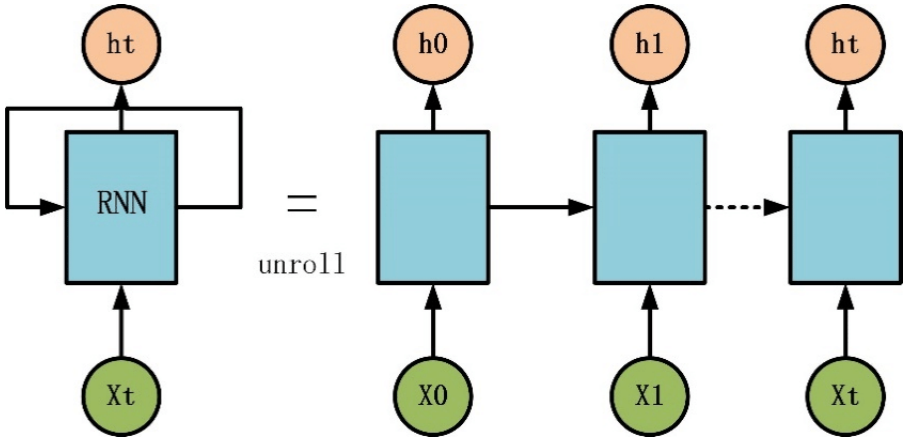

Figure 1: RNN

The problem with RNN:

RNNs are good for short predictions like for short sentence completion. But if we have a long contest, and RNN does not remember much of previous data. In other words, it requires a bunch of historical information to get right answers where RNN fail. Theoretically they should be able to do it but apparently they fail. So here comes LSTMs in the picture.

2.3.1 Long Short Term Memory (LSTM)

LSTMs is from 1990s which stands for long short -term memory. To overcome the limitations of RNN it is designed which are –

1. Complex training
2. Gradient vanishing and exploding – meaning the Gradients in the Gradient decent either go to zero or they explode all the way to infinity or maximum one
3. To over come processing long sequences difficulty

LSTM is the solution of vanishing gradient problem of RNN. The architecture of the LSTM is same as that of RNN but with a slight difference. The LSTM state has cell state, which is not there in RNN architecture. An LSTM has three inputs cell state c_{t-1}, hidden information/activation state h_{t-1} and current time input x_t.

Each LSTM cell consists of the listed.The forget gate is responsible for dropping/forgetting or retaining a value. The forget gate has an activation function of sigmoid, σ. As in figure 2, let's say that the LSTM has an input h_{t-1}, x_t and c_{t-1}. The activation function will give an output f_t, by calculation with different weights to input h_{t-1} and x_t.

The output f_t is given by:

$$f_t = \sigma[(w_{fh}{}^*h_{t-1}) + (w_{fx}{}^*x_t) + b_f] \tag{1}$$

Where w_{fh} and w_{fx} are the weights given to the h_{t-1} and x_t respectively and b_f is the bias. The f_t and c_{t-1} dot product will give c_t i.e.

$$c_{tf} = c_{t-1}{}^*f_t \tag{2}$$

The dot product should work for the same dimensionality. For example, let's say c_{t-1} has [1,4,2] and f_t has [1,0,1], then the dot product c_t is [1,0,2]. In this way it has forget the '4'. In other words, the sigmoid function restricts the output between 0 and 1. So it will, if it has to discard the previous memory it will give a output a vector of all the values which are close to zero and when it is multiplied with previous cell memory state, the multiplication will of course be zero because we have discarded the previous memory. This is how the forget gate works.

The input gate consists of two blocks i.e. Input gate and Input node. The Input gate has a sigmoid activation and Input node has tangent hyperbolic activation function, tanh activation.The input gate takes input h_{t-1} and x_t and give i_t as output and similarly input node takes two inputs h_{t-1} and x_t and gives g_t as output as shown in figure 2:

$$i_t = \sigma[(w_{ih}{}^*h_{t-1}) + (w_{ix}{}^*x_t) + b_i] \tag{3}$$

$$y_t = \tanh[(w_{gh}{}^*h_{t-1}) + (w_{gx}{}^*x_t) + b_j] \tag{4}$$

Then both i_t and g_t converges to give a dot product c_{ti}, i.e.,

$$c_t = i_t{}^* g_t \tag{5}$$

The c_{tf} and c_{ti} sums up to give the cell state c_t.

$$c_t = c_{tf} + c_{ti} \tag{6}$$

The output gate gives an output o_t as given:

$$o_t = \sigma[(w_{oh}{}^* h_{t-1}) + (w_{bx}{}^* x_t) + b_o] \tag{7}$$

The final h_t is given by the dot product of the c_t with tanh activation and o_t i.e.

$$h_t = (\tanh(c_t)^* o_t) \tag{8}$$

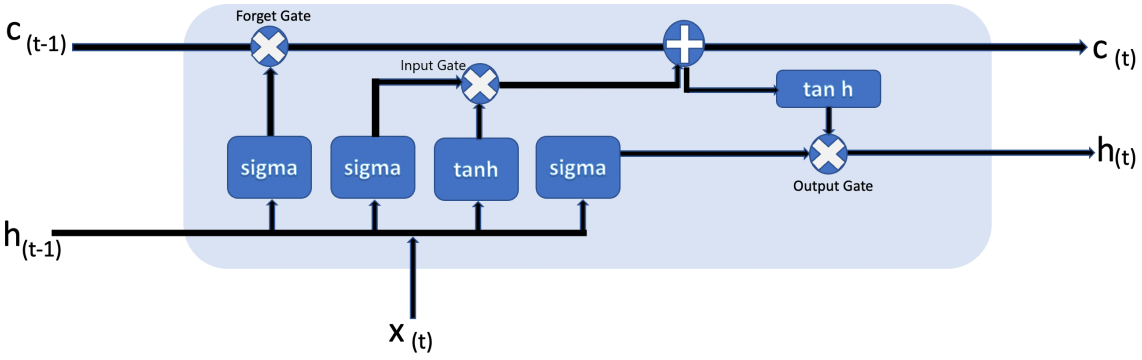

Figure 2: LSTM

2.3.2 Bidirectional Long Short Term Memory (BiLSTM)

Simple RNN which works in a single direction, that is from left to right direction. Then a particular entity will have influence from only the previous entity. We have only back propagation LSTM, which while processing the input, enables it to only access the prior information in the sequence. Recurrent networks that are bidirectional i.e., BiLSTM are actually two separate RNNs combined as shown in Figure 3. The networks may receive both forward and backward sequence information, thanks to this architecture at each time step. When using bidirectional LSTM, the inputs to the network will be processed in two different directions: one from the present to the future and the other from the future to the present. This methods differs from unidirectional in that information from future is preserved in LSTM which runs backward, and by combining the two hidden states, which can preserve data from both the present and the future at any given time. Because they can better understand the context, BiLSTMs produce excellent outcomes.

A reverse LSTM is added in a BiLSTM, in contrast to unidirectional one. Cells in this can simultaneously get context information,reverse LSTM inverts data, hidden layer combines the information from forward and reverse.Just like calculation in forward LSTM, the LSTM reverse layer, the calculation is same only the difference is the reverse direction to get the serial time information.

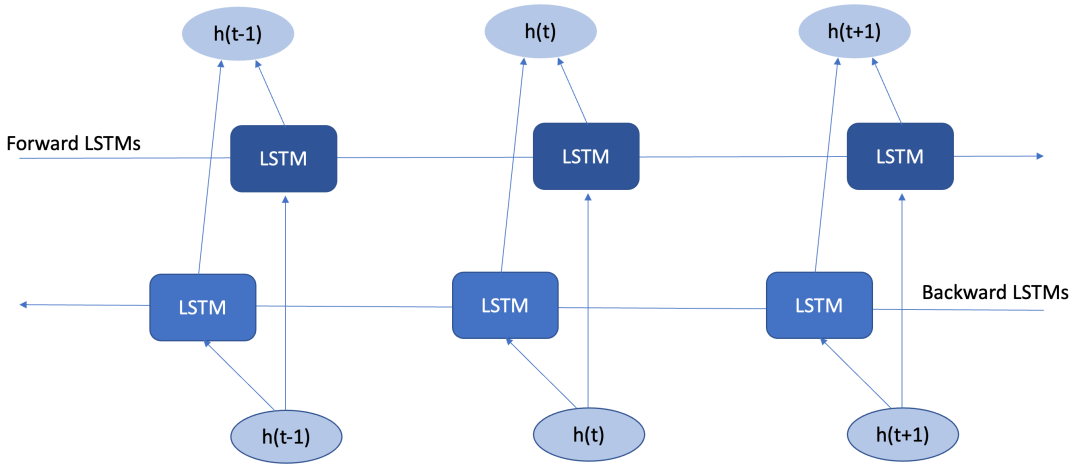

Figure 3: BiLSTM

2.3.3 Attention Bidirectional Long-Short Term Memory(Attention-BiLSTM)

A LSTM or BiLSTM will perform better with the addition of an Attention layer, which also aids in creating predictions or forecasts in an accurate sequence.Humans don't learn everything from scratch; rather, they make conclusions about new information by relating old information to it.For instance, if someone already knows how to cycle and learns how to ride a motorcycle, they won't have to learn about braking or any other fundamental concepts because they are already familiar with them.With older or regular information they add extra informations.Regular neural networks can not do this and hence it is a shortcoming of them. This can be done by the RNNs. LSTMs are special type of RNNs which are networks containing a variety of loops to preserve the data.These networkss are very useful in dealing time series as well as NLP data.As discussed earlier, there are three main types of LSTM networks:

1. forward LSTM
2. Backward LSTM
3. Bi-directional LSTM

As per the name both the forward and backward LSTM has unidirectional information processing, whereas the bidirectional LSTM can do that in both directions.And attention-BiLSTM are used to increase the performance if the BiLSTMs.

One of the most important developments in deep learning model construction in recent years is the attention mechanism. It has been widely applied to NLP issues.Consider a scenario where we must identify a person from a photo of a select group of well-known individuals. Basically, it's a picture of everyone we know in a group. In order for our awareness to help our subconscious mind, we must now recognise just one person. The mind will create an image of that individual, which we may identify by matching. This implies that our minds are exclusively focused on the artificially created image of that person. Therefore, concentrating just on one individual among a bunch might be seen as attentiveness.

The foundational LSTM or RNN model was built on an encoder-decoder architecture prior to the development of the attention mechanism. When data is processed for encoding into a context vector and a good overview of

the input data is produced, here is where encoding is applied. The remainder of this overview then covers the decoding phase, during which the model interprets and comprehends the data.If this overview is not good, then though the accuracy of the model will be good for the base model,but in the long information processing the model will give bad results.This is the long-range dependency problem of tradition recurrent networks.

Therefore, if this kind of circumstance arises, the encoder stage needs to be looking for the most pertinent information; this concept is known as "Attention."In order for the decoder state to translate it better and the model to forecast it better, the encoder states read the sequential input and summarise it. The summarised component then receives weighting from the model attention layer.

A typical attention-BiLSTM has main parts-input layer, BiLSTM layer (Forward and Backward LSTM),followed by an attention layer,fully connected and then output layer.Following input, model sends the series data to forward and reverse LSTM hidden layers, which work together to produce the processed vector. The weight vector is calculated by the Attention layer using LSTM input data. The weight vector is then combined with shallow output to create new vector input for the fully connected layer. The projected value is then calculated by the fully linked layer.A typical model of Attention-BiLSTM is as shown in figure 4.

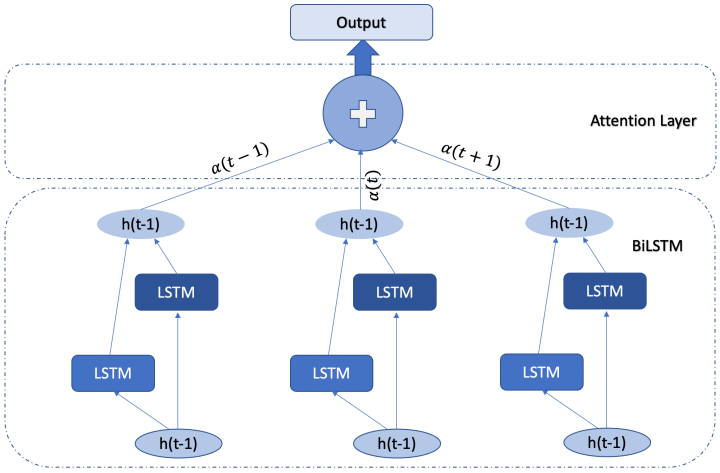

Figure 4: Attention-BiLSTM

2.4 Regression

On the premise that variable that is forecasting has a cause-and-effect relation with one or more other variables, exploratory or casual forecasting approaches are based.These techniques aid in illuminating how the magnitude of one variable affects the magnitude of another.For example, Regression analysis can be used to create an equation that demonstrates how two variables are related, for as how advertising expenditures affect the amount of sales for several products.Advertising expenses would not be taken into account if a time series approach was used to create the prediction; instead, time series method will only use historical sales data to create the forecast.

A time series variable is forecasted using econometric models, often known as regression-based or casual models, employing regression and other explanatory time series variables.For instance, a business may use a causal model to predict future sales on its amount of advertising, the income level of the population, the interest rate, and perhaps other factors. Regression analysis with time series variables resembles regression analysis in several ways.

2.4.1 Multiple Linear Regression

One of the most popular statistical methods for forecasting is the regression method.Regression models are able to describe the connection between the independent or forecasting variable, here, UPDRS Score with other independent variables here the speech features.Regression analysis is examining the connection between a continuous independent variable y and one or more other independent variables x1 through xk.Regression analysis's objective is to find a function that best captures the relationship between these variables so that a range of values for the independent variables is used for value predictions of dependent ones. The UPDRS Score is determined using the multiple linear regression approach in terms of an explanatory (independent) variable like the speech features.The forecasting using the multiple variable linear regression is given as shown in equation:

$$Y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{9}$$

Here, Y_t is the forecasted variable at time t.In this case, it is the UPDRS Score at time t. The independent speech variables are the variables $x_1, x_2, ..., x_k$.β_j parameters are the regression parameters with respect to the independent variables.This is the typical impact of a one increment in x_j on Y while keeping the other predictors constant.And the ϵ is the error term here.The value of the β_j is calculated by optimizing the sum of squared residuals, RSS.Where $\hat{Y}$ is predicted forecast and Y is actual outcome.

$$\text{RSS} = \sum(\hat{Y}\text{-}Y)^2 \tag{10}$$

2.4.2 Ridge Regression

One problem that might get encountered in traditional linear regression is that the multicollinearity can be an issue if the predictor variables are correlated highly . This may result in the model's coefficient estimations having a high variance and being untrustworthy.To overcome this problem that too without removing the independent predictors is by using the method called ridge regression.Along with the RSS term we have the another term called shrinkage penalty. as shown in equation:

$$\text{Ridge} = \sum(\hat{Y}\text{-}Y)^2 + \lambda \sum(\beta_j)^2 \tag{11}$$

When λ is equals to zero, then the regression will be same as the traditional least squares or traditional linear regression model.However, when the λ tends to infinity, the estimates of the ridge regression coefficients decrease as the shrinkage penalty gains strength.The bias-variance trade-off is where ridge regression differs from least squares regression.The fundamental principle of ridge regression is to add a small amount of bias in order to significantly reduce the variance, which results in a lower total MSE.

3 Previous Works

Numerous studies were carried out in PD to maintain and enhance the quality of life of the patients. A critical issue from a medical standpoint is the accurate identification of PD symptoms and indications at an early stage, which has roused the interest of multidisciplinary researchers to thoroughly investigate this subject. Even though PD cannot be treated permanently, the progression of the disease can be slowed down with the help of correct medication and care[9]. Routine examinations are necessary for patients to track the evolution of their disease. Since PD frequently affects patients' ability to use their motor function, the dysfunction on speech and movement can be documented and recorded employing electronic devices[3]. Various body part's motor abilities have been gathered in prior work, including typing patterns, gait abnormalities in diverse walking styles, hand motions, doing other daily tasks, and speaking patterns[4],[5],[11]. However, in addition to speech data, other signs must be captured using specific acceleration sensors mounted to wearable devices. While a non-intrusive equipment, such as a phone, can be used to capture speech. The motivation behind the current work is the ease of using these speech features that are recorded using smartphones and stored in database, for forecasting a -PD episode in the future. This straightforward approach is appropriate for achieving the need for doctors to continuously monitor the course of PD symptoms so they can act right away about medication dosing, drug side effects, and requests for additional testing to enable patients to function at their optimum.

Different statistical models as well as machine learning methods have been used to various PD dataset types. Three different forms of learning algorithms were used to gauge the degree of motor abilities. In comparison to hidden Markov models and support vector machines, the experimental results demonstrate that the neural network produced the best outcomes [15]. Forty features of speech from PD patients were retrieved using a speech dataset, and the created DNN and regression algorithm was able to identify and categorize four severity categories [8]. The study demonstrates that DNN consistently outperforms alternative machine learning classifiers in the diagnosis of Parkinson's disease [14]. DNN is capable of categorizing unstructured data, including voice and audio signals [13].

Time series forecasting uses temporal measurements from earlier observed data to mathematically estimate certain future values. Using mathematical and statistical techniques, the model was created based on particular as-

sumptions about the variable behavior of the underlying system. Time series forecasting models are used in medical applications to forecast illness development, calculate mortality rates, and evaluate potential risks across time. Like it is used for cardiovascular disease monitoring [10] and in case of kidney disease [7].

4 Methods

4.1 Parkinson's Dataset

42 participants, including 14 women and 28 men, contributed a total of 5875 recordings to the dataset used in this study.The data set can be found in the UCI machine learning repository [2].There are roughly 200 voice recordings for each patient. A recorded voice contains 16 vocal characteristics.Additionally, the dataset includes Motor-UPDRS and Total-UPDRS scoring.A common scale used by doctors to diagnose Parkinson's disease and track its progression is called the UPDRS.Total-UPDRS runs from 0 to 176, whereas Motor-UPDRS extends from 0 to 108, with 0 representing a healthy state and 108 signifying severe motor impairment, respectively. Prime features in the dateaset is as shown in table 1.

4.2 Data processing

Data mining's foundational task is the processing of data standardisation. Dimension units for various evaluation indicators can vary.The outcomes of data analysis will be impacted by such circumstances, making it necessary to remove the dimension between indicators.It is necessary to standardise data to address the issue of data metrics' comparability. Each indication is in the same order of magnitude after the original data is processed through data standardisation, making it appropriate for thorough comparison evaluation.As a result,StandardScaler normalisation is first used to handle the data before training and verification.StandardScaler is a tool used in machine learning to scale the value distribution such that the observed values' mean is 0 and their standard deviation is 1.

Features	Description
Jitter(%) Jitter(Abs) Jitter:RAP Jitter:PPQ5 Jitter:DDP	Several measures of variation in fundamental frequency
Shimmer Shimmer(dB) Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA	Several measures of variation in amplitude
NHR,HNR	Noise to tone ratio in voice
RPDE	A nonlinear dynamical complexity measure
DFA	Signal fractal scaling exponent
PPE	A nonlinear measure of fundamental frequency variation
motor_UPDRS	Clinician's motor UPDRS score, linearly interpolated
total_UPDRS	Clinician's total UPDRS score, linearly interpolated

Table 1: Features in the dataset

4.3 Attention-BiLSTM architecture

The experimental platform is Google Colab with GPU server. The model architecture is shown in figure.Important information can be both remembered and forgotten by the LSTM layer. Theoretically, model fits nonlinear data more accurately the more layers there are. But wearing too many layers can make you overdressed and take a lot of time. The model consists of a BiLSTM layer,with 120 neurons in LSTMs. This is followed by an attention layer.The attention has two functions that are build() and call(). The build() function will define the bias and weights.If the input shape is (None,7,240), then the output of the function will give a shape (240,1). The call() function will multiply the weights and add bias. Then, the tanh function is followed by softmax layer. So, if the input to the model is 3D, then the attention layer will give the output as 3D.This comprises of the attention-BiLSTM. The following layer is of another BiLSTM, which consists of 70 neurons LSTMs.This is finally connected to the Dense layer as shown in figure 5

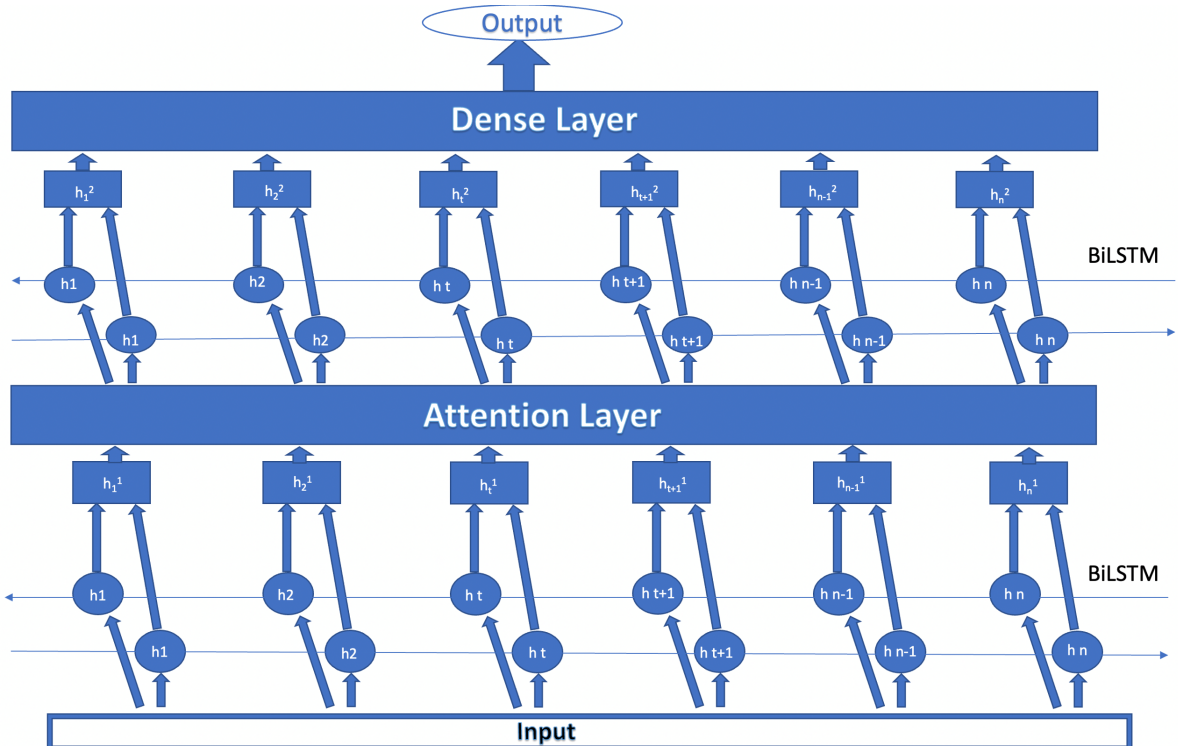

Figure 5: Attention-BiLSTM architecture

4.4 Evaluation metrics

The performance metrics used in this study are mean absolute error (MAE), mean square error (mse) and root mean square error (rmse). After building a model, we want to measure how accurate the prediction is made by the model.MAE is used for quality assessment and summarizing the model.The error here is the difference of predicted and that of actual data. The absolute value of this error is taken into account.i.e.|error|.Finally, the sum average of all the recorded absolute errors are calculated.

MSE is performance metrics that is used for predicting the accuracy of the model in this study.It is used to first calculate the difference between the predicted and actual value. The square of the difference is taken and the mean sum of the squared value is calculated finally.

RMSE is another metrics that is used in this study.According to the literature, time series should be represented using the RMSE and MAE generally. However, RMSE was widely utilised in earlier research.Forecast errors are given a lot of weight in RMSE.The square root of the mean average of the error difference of predicted and actual value is calculated in RMSE.

5 Result

The result of the study is as shown in table 2. The proposed model MAE, MSE and RMSE value is 0.0436, 0.2088, and 0.0758 respectively.The MAE value is 24.35% less than LSTM,7.01% less than BiLSTM,23.2% less than linear regression and 19.78% less than the ridge regression model.The MSE value is 23.77% less than LSTM, 12.97% less than BiLSTM, 5.58% less than linear regression and 2.24% less than the ridge regression model. The RMSE value of the proposed model is 12.67% less than LSTM, 6.70% less than BiLSTM, 2.79% less than linear regression, and 1.08% less than ridge model. The overall performance of all the models is shown in table 2.

S.No.	Models	MAE	MSE	RMSE
1	Linear Regression	0.0987	0.04618	0.2148
2	Ridge Regression	0.0945	0.0446	0.2111
3	LSTM	0.1002	0.0572	0.2391
4	BiLSTM	0.0815	0.0501	0.2238
5	Attention-BiLSTM	0.0758	0.0436	0.2088

Table 2: Result for MAE,MSE and RMSE for different models- LSTM, BiLSTM, Attention-BiLSTM, Linear Regression, Ridge Regression

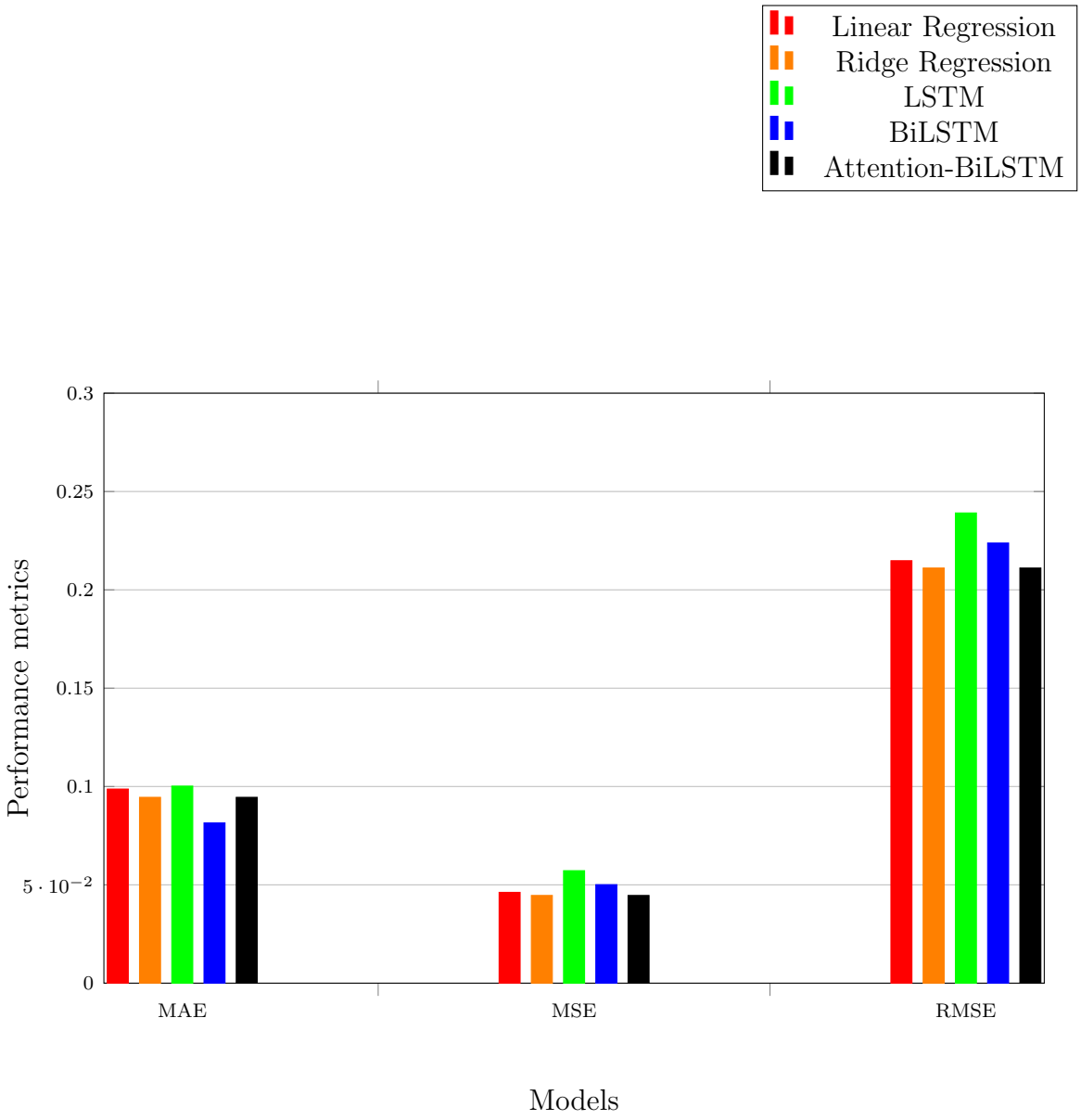

Figure 6: Performance metrics of models

In light of this, if deep learning algorithms is applied to a particular dataset, we construct a model that can accept input and produce output.Now, we employ a metric called as loss to evaluate a model's performance.This loss specifically measures the error that the model causes.A high loss number typically implies that model is giving inaccurate output, whereas a low loss value denotes that the model has less flaws.Additionally, a cost function is typically used to quantify the loss and it measures the mistake in several ways.The problem being addressed and data being used are typically determinants of the cost function selected. For binary classification issues, for instance, cross-entropy is frequently used.It calculates model's error on training set.On the other hand, a deep learning model's performance on the validation set is evaluated using a statistic called validation loss.A piece of the dataset designated as the validation set is used to verify the model's efficacy.Furthermore, the validation loss is calculated at the end of each period. This tells us whether the model requires to be adjusted or tuned further.The training loss and validation loss are typically combined on a graph in most deep learning applications. This is done to evaluate the model's performance and determine what needs to be tuned. The training-validation loss curves of Attention-BiLSTM is as shown in figure 7.

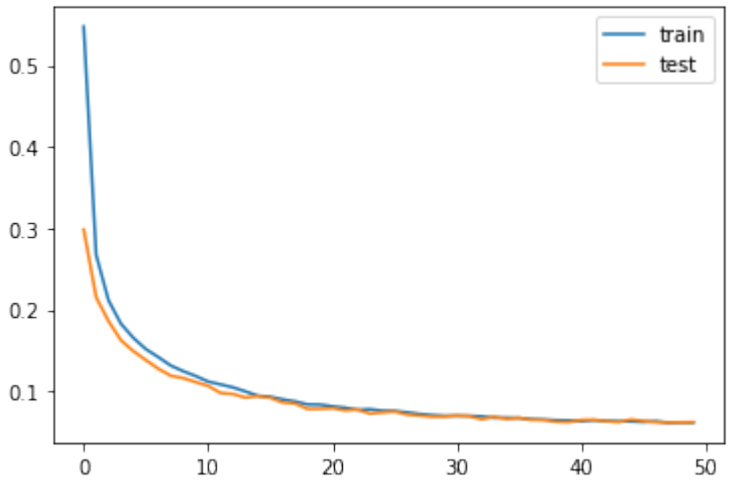

Figure 7: Attention-BiLSTM Training-validation loss curve

The predicted output of UPDRS score against the actual data points of the forecast is as shown in figure 8 for the proposed model.The predicted output from a multivariate linear regression and ridge regression is also shown in figure 9.

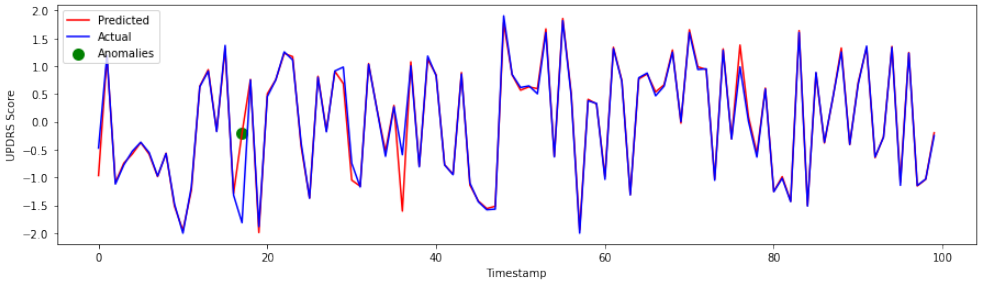

Figure 8: Attention-BiLSTM Forecast

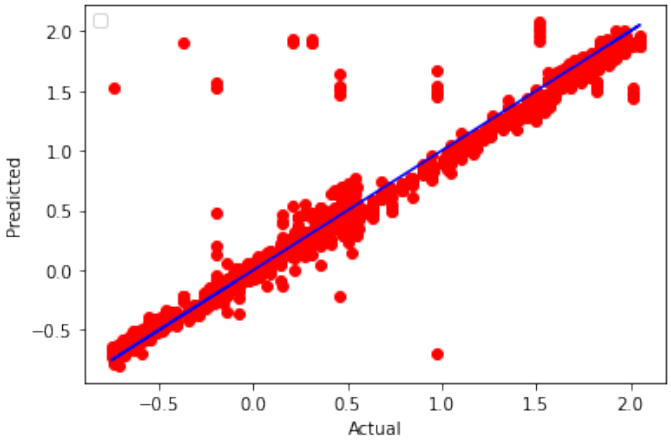

Figure 9: Linear Regression Forecast.Where red dots represents predicted value of the score and blue line represents the actual outcome.

6 Conclusion

Before a consultation with a clinician, remote monitoring of UPDRS using voice measures is an effective screening step.The creation of computational tools employing DNN approaches can help the medical expert predict the patient's PD progression more quickly and identify the subjects earlier.Following the development of clinical PD symptoms regularly can be a good guide for clinical diagnosis.Although diagnosing PD is frequently difficult and tracking the progression of the symptoms takes time, even in the early stages, slight vocal differences may be easily discernible.With this knowledge, voice recordings from future patients can be used to monitor PD. A Bi-LSTM forecasting model based on the Attention mechanism is created in this work.Dataset of speech features i.e 16 speech features as well as UPDRS score is used for the forecasting here. The dataset is downloaded directly from the UCI machine learning repository.The data is preprocessed before feeding to a model. In this work Standard scaling is done, so that we have all the features in the same scale and this also increases the performance accuracy i.e. helps the model learn fast.An attention layer is used in a BiLSTM model(LSTM with 120 neurons) followed by another BiLSTM model(LSTM with 70 neurons).Evaluation metics used are MAE, MSE and RMSE. Which shows a result of value is 0.0436, 0.2088, and 0.0758 for MAE,MSE and RMSE respectively.The proposed model shows a better performance as compared to it's counterpart models in this study.It is proven that the Attention mechanism has a beneficial impact on forecasting accuracy.

7 Future Scope

The particular study can be extended by using different architecture or hyperparameter tuning to further increase the performance accuracy of forecasting. Different performance metrics can also be employed to further understand the model performance.A larger dataset containing more data can be run on a forecasting model.A different combination of features can be used for forecasting the future value of a particular feature.

8 References

1. Tysnes, O. B. & Storstein, A. Epidemiology of Parkinson's disease. J. Neural Transm. 124, 901–905 (2017).

2. ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.

3. Sarbagya Ratna Shakya, Chaoyang Zhang, and Zhaoxian Zhou. International Journal of Machine Learning and Computing, Vol. 8, No. 6, December 2018.

4. Teresa Arroyo-Gallego, María Jesus Ledesma-Carbayo, Álvaro Sánchez-Ferro, IanButterworth, Carlos S Mendoza, Michele Matarazzo, Paloma Montero, RobertoLópez-Blanco, Veronica Puertas-Martin, Rocio Trincado, et al.2017. Detection ofmotor impairment in Parkinson's disease via mobile touchscreen typing. IEEETransactions on Biomedical Engineering 64, 9 (2017), 1994–2002.

5. Shimon Sapir, Lorraine Ramig, and Cynthia Fox. 2008. Speech and swallowing disorders in Parkinson disease. Current opinion in otolaryngology & head andneck surgery 16, 3 (2008), 205–210.

6. K Bache and M Lichman. 2013. UCI Machine Learning Repository. University ofCalifornia, School of Information and Computer Science, Irvine, CA (2013).

7. Rajesh Ranganath Jamie S. Hirsch David Blei Perotte, Adler and NoÃĬmie Elhadad.2015. Risk prediction for chronic kidney disease progression using heterogeneouselectronic health record data and time series analysis. AMIA Annual SymposiumProceedings 22, 4, 872–880.

8. Elmehdi Benmalek, Jamal Elmhamdi, and Abdelilah Jilbab. 2015. UPDRS tracking using linear regression and neural network for Parkinson's disease prediction. International Journal of Emerging Trends & Technology in Computer Science(IJETTCS) 4, 6 (2015), 189–193.

9. Mehrbakhsh Nilashi, Othman Ibrahim, and Ali Ahani. 2016. Accuracy improvement for predicting Parkinson's disease progression. Scientific reports 6 (2016),34181.

10. C Bui, N Pham, A Vo, A Tran, A Nguyen, and T Le. 2017. Time Series Forecasting for Healthcare Diagnosis and Prognostics with the Focus on Cardiovascular Diseases. In International Conference on the Development of Biomedical Engineering in Vietnam. Springer, 809–818.

11. W Nanhoe-Mahabier, AH Snijders, A Delval, V Weerdesteyn, Jaak Duysens, SOvereem, and BR Bloem. 2011. Walking patterns in Parkinson's disease with and without freezing of gait. Neuroscience 182 (2011), 217–224.

12. Chien-Wen Cho, Wen-Hung Chao, Sheng-Huang Lin, and You-Yin Chen. 2009. A vision-based analysis system for gait recognition in patients with Parkinson's disease. Expert Systems with applications 36, 3 (2009), 7033–7039.

13. Srishti Grover, Saloni Bhartia, Abhilasha Yadav, KR Seeja, et al.2018. Predicting Severity Of Parkinson's Disease Using Deep Learning. Procedia computer science 132 (2018), 1788–1794.

14. Resul Das. 2010. A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications 37, 2 (2010), 1568–1572.

15. Bryan T Cole, Serge H Roy, Carlo J De Luca, and S Hamid Nawab.2014. Dynamical learning and tracking of tremor and dyskinesia from wearable sensors. IEEE Transactions on Neural Systems and Rehabilitation Engineering 22, 5 (2014), 982–991.

16. Parkinson's Disease Foundation. (n.d.). Retrieved from http://www.pdf.org/

17. Bishop, C. M., Pattern Recognition and Machine Learning, Springer Science Business Media, 2006.

18. Chiuchisan, I. and Geman, O., An Approach of a Decision Support and Home Monitoring System for Patients with Neurological Disorders using Internet of Things Concepts, WSEAS Trans. Syst., 13, pp: 460-69, 2014.

19. Lipton, Z.C., Berkowitz, J. and Elkan, C., 2015. A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019.

20. Patel, S., et al., Monitoring Motor Fluctuations in Patients With Parkinson's Disease Using Wearable Sensors, IEEE Trans. Inf. Technol. Biomed., 13(6), pp: 864-873, 2009.

21. National Parkinson Foundation. (n.d.) Retrieved from http://www.parkinson.org.

22. Zeng, Yang Y A, Feng H A, etal. A convolution BiLSTM neural network model for Chinese event extraction[J]. Natural Language Understanding and Intelligent Applications, 2016(-):275-287.

23. Wang, Huang Y A, Zhao M A, etal. Attention-based LSTM for aspect-level sentiment classification[J]. Proceedings of the 2016 conference on empirical methods in natural language processing, 2016(-):606-615.

24. Bin, Yang Y A, Shen Y A, etal. Describing video with attention-based bidirectional LSTM[J]. IEEE transactions on cybernetics, 2018, 7(49):2631-2641.

25. Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Eleventh annual conference of the international speech communication association. 2010.

26. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 1998, 6(02): 107-116.

27. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735- 1780.

28. Wang, Y., Huang, M., & Zhao, L. (2016, November). Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 606-615).

29. Zhang, Peter G. Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputing, 2003(50):159-175.

30. Ismail, N.,H., Du, M., Martinez, D., & He, Z. (2019,September). Multivariate Multi-step Deep Learning Time Series Approach in Forecasting Parkinson's Disease Future Severity Progression.In 10th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics in Niagara Falls, NY.10.1145/3307339.3342185.