\section{Data Processing}~\label{sec:processing}

The majority of today's businesses rely on data to make key choices. However, we have a plethora of data at our disposal. Using new breakthrough technologies like Artificial Intelligence (AI) and Deep Learning for smart decision making and driving corporate success is of no use, if the correct data processing procedures aren't used.

Data processing is the process of changing data from one format to another that is more useable and desirable. Each dataset is unique and specific to the project's needs. The retrieved data is converted into a format that the algorithms can interpret.

Data is required to train deep learning models, A lot of data. Unfortunately, most data are messy, and our models are quite sensitive to this. As a result, we must be cautious when preparing our data in order to acquire the greatest outcomes.

\begin{figure}[h]

\centering

\includegraphics[width=0.8\textwidth]{Data\_processing.jpg}

\caption{Overview of data processing}

\label{fig:Data\_processing}

\end{figure}

\subsection{Feature Extraction}~\label{subsec:featureextraction}

Huge data sets include a large number of variables to process, which demands a lot of computational resources. Feature extraction refers to the strategies for selecting and/or combining variables into features in order to reduce the quantity of data that has to be processed while still properly and thoroughly representing the original data set ~\cite{haq2021forecasting}. It decreases the number of processing resources without sacrificing crucial or relevant data. In this paper ~\cite{kelotra2020stock} the feature extraction procedure uses data from the live stock market to extract the essential feature based on technical indicators such as double exponential moving average (DEMA), rate of change (ROCR), average true range (ATR), relative strength index (RSI), simple moving average (SMA), average directional index (ADX),commodity channel index (CCI), moving average convergence divergence (MACD), on balance volume (OBV), Parabolic SAR (SAR), exponential moving average (EMA),and triple exponential moving average (T3). ~\cite{LONG2020106205} Convolutional Neural Network (CNN) based feature extraction techniques is used to find the significant similarities and variances in trading patterns are found. In addition, pooling processes eliminates non-essential trading features. Dropout procedures helps to solve the problem of over-fitting in feature extraction and enhance the model's generalisation.

Due to its feature extraction ability ~\cite{hao2020predicting} CNN model is used to extract multiple time scale features, and the output from CNN models is referred to as a feature map. The short-term features is immediately viewed as the daily price series. Two layer CNN is utilised in this study to extract medium-term and long-term features, with the output from the one-layer CNN being considered medium-term and the output from the two-layer CNN being considered long-term features. ~\cite{mohanty2021financial} Deep learning algorithms, such as Auto Encoder, offer a greater feature extraction capacity. Several AEs are stacked one on top of the other in this study, the first layer is trained, and the current AE's hidden layer is provided as input to the next layer. The SAE technique is used to accomplish feature extraction, with the AE acting as the feature extractor for several neural networks.

\subsection{Clustering of Features}~\label{subsec:clustering}

Irrelevant and redundant features increase the data's dimensionality. When data has a large dimensionality, it is computationally expensive and reduces the accuracy of deep learning algorithms. Clustering algorithms clusters data together based on their similarity. It elucidates key features in the data that can aid in improving the model's performance. Feature clustering enhances learning algorithms' effectiveness, and it has been shown that cluster analysis is more successful than typical feature selection techniques.

~\cite{kelotra2020stock} Highly significant features are needed to be sent to the proposed model Rider-MBO-based-Deep ConvLSTM model, in order to do so, features which are obtained from the technical indicators are clustered using Sparse-Fuzzy C-Means (FCM) algorithms.

Since the investors trading behaviours are considered in this paper ~\cite{LONG2020106205}, clustering algorithm is used to aggregate the investors to consider their trading indicators. Based on the investor clustering profile, K-means algorithm is applied.

\subsection{Correlation of Stocks}~\label{subsec:correlation}

Correlation is a statistical measure of how assets, such as stocks, move in relation to one another. Correlation can be utilised to get insight into the wider market's general nature. If the correlation is high, it means that the stocks move in a lockstep manner. On a scale of -1 to +1, correlation is assessed. A result of +1 indicates a complete positive correlation between two assets. The reading of a complete negative correlation is -1. It's uncommon to find perfect positive or negative relationships.

It is hard to forecast stock market by only using target stock's historical data, thus by combining market data from relevant stocks, we can depict the market influence on the target stock. ~\cite{LONG2020106205} In this paper, by creating companies knowledge graph for the relevant stocks to the target and by applying node2vec algorithm correlation between the stocks are determined. The cosine similarities are then computed, and stocks that have top-K similarities to the target stock are considered relevant. Those relevant stocks historical data are also used to forecast the target stock.

\subsection{Frequency Decomposition}~\label{subsec:decomposition}

Because of the enormous number of unexpected market movements and trade noises, financial data has a complicated structure of irregularities and roughness. These irregularities can cause incorrect accuracy for the prediction model. Certain time series data are so variable and stochastic, analysing and forecasting them are difficult. As a result, in order to improve the prediction models, the prediction models require a denoising method to clean the data. Frequency decomposition makes data processing much easier and reduces the complexity of data.~\cite{HadiRezaei:2021a} EMD and CEEMD are used as frequency decomposition methods, which decomposes the sequential data into several frequency spectra. CEEMD can be more successful than the EMD algorithm when it comes to training deep neural networks, because the CEEMD algorithm has improved over the EMD algorithm. EWT is utilized to decompose raw time series data into many sub-layers in this work ~\cite{liu2020improved}. Each sub-layer are further divided into Training set 1 and testing set 1. The LSTM model is trained using training set 1 and tested with testing set 1, yielding a predicting result.

\subsection{Dimensionality Reduction}~\label{subsec:post-processing}

Large quantity of data may contain redundancy and unnecessary information, predicting with the original data may result in severe inaccuracies. Dimensionality reduction is used as a data pre-processing approach in stock market prediction to prevent the problem of over-fitting. ~\cite{lee2020stock} Embedded layer and Automatic encoders are used to reduce high-dimensional data into low-dimensional data. The embedded layer in the ELSTM model uses matrix transformation, were as an automatic encoder is employed in conjunction with a continuous restricted Boltzmann machine (CRBM) to minimise the size of the stock vector in the AELSTM model. Apart from that to avoid the overfitting problem, random column-wise shuffling (Data augmentation approach) is utilised, and trend sampling, a mini-batch sampling method, is used to emphasise current market trends.

\subsection{Data Post-processing}~\label{subsec:post-processing}

When it comes to data post-processing, error modelling is a popular strategy, with many research focusing on the vector error correction. Because of its robustness and performance, the outlier robust extreme learning machine (ORELM) model may be a viable candidate for data post-processing. In this study, data preparation and postprocessing methods EWT and ORELM are employed, with EWT based decomposition and ORELM based error correction. For each sub-layer, the ORELM model is employed for error modelling based on past predicting results. It is used to overcome the problem of low-frequency component prediction accuracy.