IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304393 |

A Machine Learning for Prediction to Order Book Stock Pricing

Mrs. K. Gandhimathi, Noah Prashanth I, Prakash S, Sriprakash P

Assistant Professor, Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

Department of CSE, Muthayammal Engineering College (Autonomous), Rasipuram, Tamil Nadu, India

ABSTRACT: The proposed solution is comprehensive as it includes pre-processing of the stock market dataset, utilization of multiple feature engineering techniques, combined with a customized deep learning based system for stock market price trend prediction. We conducted comprehensive evaluations on frequently used machine learning models and conclude that our proposed solution outperforms due to the comprehensive feature engineering that we built. The system achieves overall high accuracy for stock market trend prediction. With the detailed design and evaluation of prediction term lengths, feature engineering, and data pre-processing methods, this work contributes to the stock analysis research community both in the financial and technical domains.

KEYWORDS: Long short term memory, Principal component analysis, Recurrent neural networks, Artificial neural network

I. INTRODUCTION

Since the inception of the stock capital markets investors have attempted to forecast the share price movements. However, at that time, the data available to them was quite limited and the approaches to processing this data were quite simple. Since then, the amount of data available to the investors has expanded significantly and new ways of processing this data have been introduced. Currently, even with all the technical progress and advanced trading algorithms, the ability to correctly predict the stock price movements still remains an extremely challenging task for most researchers and investors. Traditional models based on fundamental analysis, technical analysis, and statistical methods, e.g., regression in [1], which were used for decades, often can not fully capture the complexity of the issue at hand. In particular, they are not suitable for the data with high cardinality, such as the Limit Order Book (LOB) data.
The recent advancement in the Machine Learning methods and proliferation of the market, fundamental and alternative data in the digital format led to numerous attempts to adopt these models for the stock price prediction task, e.g., [2,3,4]. Some researchers were already able to demonstrate quite impressive results in this area, in particular using LOB data as a main source, e.g., [5]. In this paper, the focus is on the critical evaluation of the practical usefulness of state-of-the-art Machine Learning and Deep Learning models based on LOB data for stock price predictions. A more detailed formulation of the research problem, motivation, goals, and the structure of this paper is presented in the paragraphs below.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304393 |

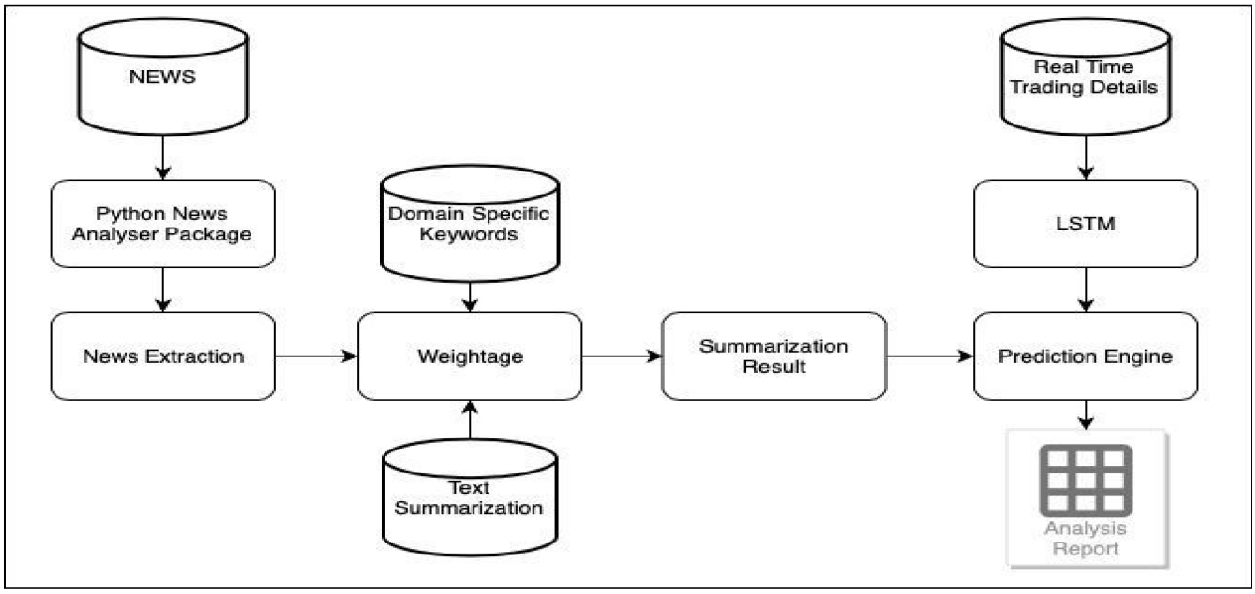

Fig 1: Stock Price Prediction Using Machine Learning

Progress in the machine and deep learning opened new opportunities for building stock price movement prediction models based on time-series data characterised by high cardinality, such as the LOB data. As a consequence this area has been the focus of increasing research interest over the recent years. Prediction performance of the suggested models is usually claimed to be rather high, for some state-of-the-art Machine Learning and Deep Learning models (e.g., [6,7]), according to the authors, accuracy is above 80%. From the practical perspective these results look too good to be actually reproducible in real world stock trading. Thus, these models require a detailed investigation, which we conduct in this paper.

This survey is focused on the critical evaluation of the current studies in the subject area of stock price movement predictions based on LOB data and identification of the improvements required and directions for further research.
In addition to this introductory section, the paper is organised into three main sections:

Section 2 contains an overview of the strategies for stock prediction based on the market data. At the beginning there is an introduction to the three core types of data used in trading: market data, fundamental data, and alternative data. The subsequent discussion will focus on market data such as stock execution prices and volumes and LOB data. Next there is an overview of the market data-based trading approaches, with their comparison, evolution trends analysis, and the respective conclusions. Among those conclusions are that the most promising data type for further analysis is LOB and the most promising model classes are Machine Learning and Deep Learning.

Section 3 is focused on a critical review of the empirical research on the Benchmark LOB dataset. At the beginning of this section there is a detailed description of the benchmark LOB dataset with the evaluation of its merits and issues. Next, there is a subsection devoted to the comparison and critical evaluation of the Machine Learning and Deep Learning models based on the Benchmark LOB dataset. At the end of this section there is a discussion of the data processing approach, experimental setup, and results for one of the state-of-the-art models, for which the experiment was reproduced.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304393 |

II. RELATED WORK

The financial metrics, such as revenue, profit, free cash flow, etc., which define the equity value of a particular company, are considered as company-specific fundamental data. The second category is the external fundamental data, which include the macroeconomic and industrial indicators relevant to the selected company, such as GDP of the country of operations, or the iron ore price for a steel producing company. The main assumption of the trading strategies relying on fundamental data is that the stock price will converge towards its fair value defined by the above-mentioned factors. For example, Bartov et al. [2] concluded that the well-known post-earnings announcement drift phenomenon is caused by the unsophisticated investors' delayed response to the new information. This suggests that trading strategies could exploit this market inefficiency and generate excess returns by rapidly and correctly responding to the new fundamental data inflows. Similarly, a more recent study [3] proved that the trading strategy exploiting the slow reaction of oil companies' investors to the changing oil price can be profitable.

The proliferation of the internet and the production of large quantities of digital data in recent years created a highly valuable source of insights on potential stock price movements. The variety of this alternative data is unprecedented, it can be any insightful information about the company, starting from such obvious examples as announcements on a companies' websites, rumours in the news blogs and on social media about a company, and ending with much more exotic data, such as the number of new positions posted on the company's recruiting section of the website, number of the visitors in the online store, or even satellite photos of the number of cars parked near the company's store or of the fields of a grain producing company. For example, the authors of this research [4] concluded that the satellite photos of the parking lots near a company's stores provide useful information for assessing a retailer's performance. According to the authors, the trading strategies utilising this data can generate extra profit at the expense of less informed market participants, who are making their investment decisions based on the earnings announcements. This is possible since only some investors have exclusive access to this information, while others have to rely on the official financial results announcements, which happen with a substantial time lag compared to the almost real-time satellite photos collection.
Market data comprises all the trade-related statistics that can be collected from the exchanges or other trading platforms, such as the flow of the orders, stock price, and trading volume. This type of data plays a pivotal role in intra-day trading and especially in High-Frequency Trading (HFT). HFT firms generated enormous profits in recent years, which provides empirical evidence that this type of data deserves the attention of researchers. In addition, this market data is often available at an extremely fine scale. With a time series interval that can be under 1 millisecond, a reasonable number of points for analysis can be collected even for a period as short as one trading day. Further to this, the trading strategies relying on fundamental data and alternative data generally need a longer investment horizon with unpredictable duration since the period of convergence to the target price in these strategies depends on how fast other market participants will process and react to this information, which could vary substantially and could be difficult to predict.

III. METHODS

Earlier studies were using predominantly different datasets, experiment setups and even the metrics measuring the models performance, were varying widely. All these factors are making them almost incomparable. Reproducibility of many of these experiments was also poor since datasets and code were made publicly available for only a few of the studies considered. The situation was improved after the first public benchmark LOB dataset was published. This work established a common platform for the research in this area by allowing a greater standardisation of experiment setup and performance metrics in addition to the benchmark LOB dataset itself. Recent state-of-the-art studies are often using this dataset to compare the results against the other models. However, only a few authors were used their models to conduct trading simulations such as in these studies and to calculate potential profits from the strategy based on the model predictions. Thus, it is possible to assess the practical value of just a few of the model suggested.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304393 |

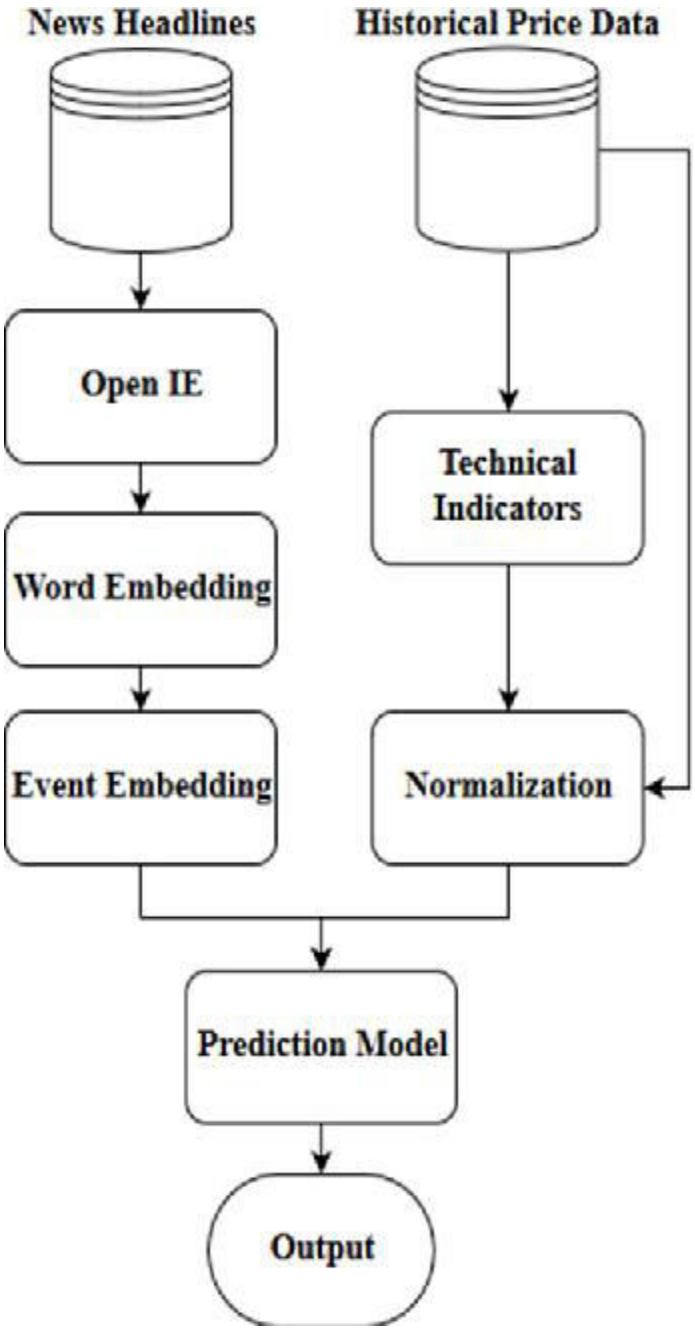

Fig 2: Stock Price Prediction Work Flow

The other problem affecting the practicality of these studies is that the transaction costs were often not taken into consideration even when this trading simulation was undertaken, with rare exceptions. The other unrealistic assumption, which is embedded in the above-mentioned benchmark LOB dataset and thus affecting all the studies using this data, is that transactions could be executed at the mid-price. The mid-price is just a simplifying approximation of the actual execution price. The former is calculated as an average of the best bid and offer prices, while the latter would be the best offer for buying and best bid for selling using market orders. This type of order would be required to ensure timely execution. It is clear that a round trip transaction in either direction with buying at best offer and selling at best bid would results in larger spread-related transaction costs than if the mid-price execution is assumed. At the same time, mid-price is still important for market-making strategies for properly positioning the bid and ask limit orders relative to the expected mid-price.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304393 |

IV. RESULT ANALYSIS

The best performance among those deep learning models was demonstrated by the DeepLOB and TransLOB architectures [7]. Each of them according to their authors demonstrating Accuracy and F1 scores well in excess of 80%. If these results are reproducible in real stock trading, these models could be potentially employed by market makers for setting their bid and ask quotes. However, we are interested in assessing whether these models could be used to generate buying and selling signals which could be incorporated in the trading strategies of active traders. Thus, the question is whether these models can be used to develop profitable arbitrage strategies. In practice there are serious concerns that this could be the case. Firstly, because of the earlier described issues with the benchmark dataset. Secondly, because of the potential flaws in the experiment setups. Thirdly, because of the ignored transaction costs and assumed mid-price execution the expected profitability of the trading strategies based on these models could be overestimated. As it would be proved later these two factors alone can make the strategies based on the suggested models unprofitable.

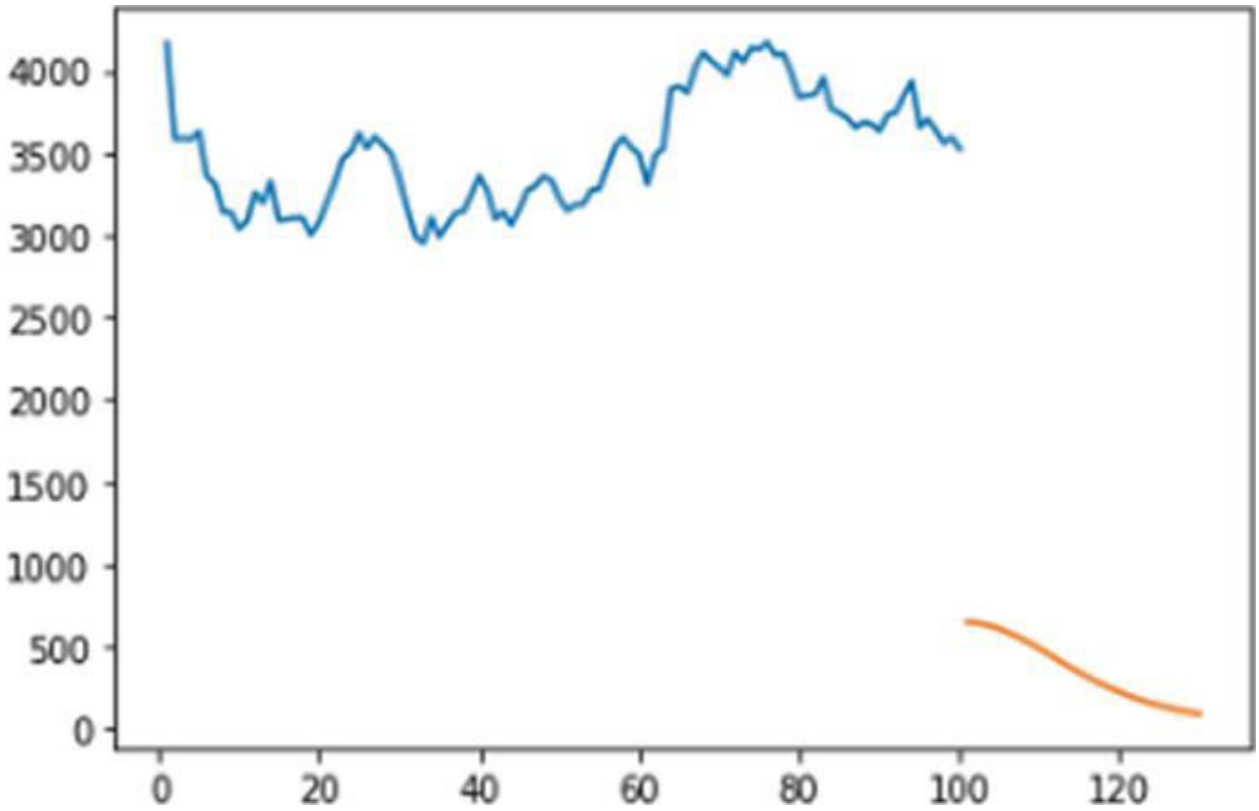

Fig 3: Result analysis

They set out a the deep neural network method with a combination of convolution layers and Long Short-Term Memory (LSTM) units DeepLOB, which was exploited to develop stock trading strategy based on the LOB data. This approach demonstrates better prediction power than any other existing algorithms relying on the LOB as a source for the feature extraction at the time of publication. Authors claimed higher F1 score of DeepLOB compared with the following models: RR, SLFN, LDA, MDA, MCSDA, MTR, WMTR, BoF, N-BoF, B(TABL), and C(TABL). The authors tested their results on the two datasets, one of them was the benchmark LOB dataset, the other was a massive one year long sample with 134 million data points based on the London Stock Exchange LOB data. A trading simulation was also conducted which demonstrated the profitability to be statistically higher than zero.

A second unrealistic assumption of the authors is that they can buy and sell stock at the mid-price. In reality execution is not guaranteed at this price. If trader wants a guaranteed execution of his order he would need to buy stock at the current best offer, which is higher than mid-price or sell at the current best bid, which is lower than the mid-price. The authors explain their mid-price approach assuming that it is possible to submit the limit order at the better price instead

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304393 |

of executing a market order. Although, it could be possible that this limit order will be executed in the desired time-frame, it is not guaranteed. Thus, mid-price assumption as well as an absence of transaction costs are potentially overestimating the profitability of the trading strategy based on the DeepLOB model.

V. CONCLUSION

The LOB as an input data for the intra-day stock price prediction has received substantial academic attention over the last decade and proved to be one of the most valuable data sources for features extraction. The number of studies and their comparability substantially increased. However, this dataset suffers from a number of issues, such as dated information, inherently unbalanced distribution between three classes, five stocks comprising this dataset are indistinguishable, potential processing issues and unrealistic mid-price execution assumption embedded in the classes labels. All these could substantially bias the results of experiments performed with this dataset. Nevertheless, a number of the Deep learning models have demonstrated a strong performance on this dataset based on standard statistical metrics such as Accuracy, Precision, Recall, and F1 score. However, further analysis has revealed that strategies based on these models can generate consistent profit only if some unrealistic conditions are assumed. One of them is the absence of transaction costs and the second is mid-price execution. Further, it was noted that some of these deep learning models are prone to over-fitting, which is limiting their generalising capability. The above-described issues in the data, models and experimental setups suggest that there is room for further research in each of these three domains.

REFERENCES

1. Chang, P.C.; Liu, C.H. A TSK type fuzzy rule based system for stock price prediction. *Expert Syst. Appl.* **2008**, *34*, 135–144. [Google Scholar] [CrossRef]
2. Bartov, E.; Radhakrishnan, S.; Krinsky, I. Investor sophistication and patterns in stock returns after earnings announcements. *Account. Rev.* **2000**, *75*, 43–63. [Google Scholar] [CrossRef]
3. Moore, J.; Velikov, M. Oil Price Exposure, Earnings Announcements, and Stock Return Predictability. Earnings Announcements, and Stock Return Predictability (20 January 2019) 2019. Available online: **https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3164353** (accessed on 10 December 2021).
4. Katona, Z.; Painter, M.; Patatoukas, P.N.; Zeng, J. On the capital market consequences of alternative data: Evidence from outer space. In Proceedings of the 9th Miami Behavioral Finance Conference, Coral Gables, FL, USA, 14–15 December 2018. [Google Scholar]
5. Kercheval, A.N.; Zhang, Y. Modelling high-frequency limit order book dynamics with support vector machines. *Quant. Financ.* **2015**, *15*, 1315–1329. [Google Scholar] [CrossRef]
6. Zhang, Z.; Zohren, S.; Roberts, S. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Trans. Signal Process.* **2019**, *67*, 3001–3012. [Google Scholar] [CrossRef] [Green Version]
7. Wallbridge, J. Transformers for limit order books. *arXiv* **2020**, arXiv:2003.00130. [Google Scholar]
8. Fu, T.C.; Chung, C.P.; Chung, F.L. Adopting genetic algorithms for technical analysis and portfolio management. *Comput. Math. Appl.* **2013**, *66*, 1743–1757. [Google Scholar] [CrossRef]
9. Zhang, Y.; Wu, L. Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network. *Expert Syst. Appl.* **2009**, *36*, 8849–8854. [Google Scholar]
10. Majhi, R.; Panda, G.; Sahoo, G.; Dash, P.K.; Das, D.P. Stock market prediction of S&P 500 and DJIA using bacterial foraging optimization technique. In Proceedings of the 2007 IEEE Congress on Evolutionary Computation, Singapore, 25–28 September 2007; pp. 2569–2575. [Google Scholar]

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details