FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Mohammed Fuseini	**Ghana**	**mfuseini133@gmail.com**	
Hakoilonga Panduleni Hamalwa	**Namibia**	**Hakoilonga@gmail.com**	
Aman Prasad	**India**	**amanprasad.cet@gmail.com**	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team member 1	**Mohammed Fuseini**
Team member 2	**Hakoilonga Panduleni Hamalwa**
Team member 3	**Aman Prasad**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
Note: You may be required to provide proof of your outreach to non-contributing members upon request.

N / A

PART 1:

Q1. Data Understanding:

The paper by Sagaceta Mejía et al. (2022**)** uses historical stock-price data to predict stock-market movements, focusing specifically on emerging markets. The authors source daily closing-price data to derive technical indicators, which serve as input variables for predictive modelling. These indicators are calculated with mathematical formulations that capture historical trends, momentum, volatility, and market sentiment. Examples include moving averages (MA), the relative strength index (RSI), moving-average convergence divergence (MACD), and Bollinger Bands (Sagaceta Mejía et al., 2022).

Technical indicators play a critical role in forecasting stock-price trends because they distill complex market behaviour into quantifiable metrics. These metrics encapsulate historical patterns and dynamics that are not immediately observable in raw price data, enabling traders and models to interpret and predict future price movements more effectively. Indicators such as RSI and MACD, for instance, help identify overbought or oversold market conditions and momentum shifts, thereby facilitating timely investment decisions and risk management (Chong & Ng, 2008).

The significance of employing such indicators stems from their ability to incorporate market psychology and investor behaviour implicitly. Because technical indicators reflect past investor actions, they offer predictive insights into future behaviour and market direction—particularly in emerging markets, which are characterised by high volatility and inefficiencies. This predictive capability is crucial for enhancing the accuracy of neural-network models, which benefit significantly from structured, signal-rich inputs for optimal performance (Sagaceta Mejía et al., 2022; Han et al., 2021).

Q2. Security Understanding

Selected Fund: IVV (iShares Core S&P 500 ETF)

The iShares Core S&P 500 ETF (IVV) is an equity exchange-traded fund (ETF) and passive index fund designed to replicate the performance of the S&P 500 Index. This index represents approximately 80% of U.S. equity market capitalization, encompassing large-cap companies across diverse sectors. IVV was launched on May 15, 2000, and has since grown substantially, boasting assets under management (AUM) of over $450 billion as of 2025. With a minimal expense ratio of just 0.03%, IVV provides investors with a cost-efficient mechanism for gaining broad market exposure.

Historically, IVV has demonstrated robust performance, reflecting the long-term upward trajectory of the U.S. stock market. At inception, the ETF was priced at around $100, appreciating to approximately $120 by 2010. Over the subsequent decade, IVV witnessed substantial growth, reaching about $320 by 2020. As of 2025, the price stands at roughly $500, translating to a compound annual growth rate (CAGR) of about 7.5%, adjusted for dividends.

The authors of the referenced paper opted for a classification approach rather than regression when predicting stock-market movements. This choice aligns closely with practical trading decisions and risk-management strategies, focusing on forecasting directional movements (up or down) instead of exact price changes. Such an approach is more actionable and appropriate given the inherent noise and volatility present in financial markets, particularly in emerging economies (Bao et al., 2017).

Alternative methods for defining classification variables

1. **Quantile-based movement**

   This approach divides returns into tertiles, labelling returns that exceed the 66th percentile as **"up,"** those between the 33rd and 66th percentiles as **"neutral,"** and those below the 33rd percentile as **"down."**

2. **Event-driven classification**

   In this method, market events—such as a significant volatility spike (e.g., a 50% increase in volatility)—determine the class label, irrespective of the direction of the price movement.

Q3. Methodology Understanding

New Section 2: Data Subcategories:

1. **Data Sources**: Emerging market stock indexes and ETFs, primarily sourced from Bloomberg and Yahoo Finance.
2. **Data Processing:** Data cleaning procedures, including identification and adjustment for outliers and corporate actions such as stock splits and dividend distributions.
3. **Feature Engineering:** Calculation of 35 technical indicators using specialized libraries like TA-Lib and custom Python scripts, designed to enhance the predictive capability of the raw financial data.
4. **Label Construction:** Creation of binary labels based on the formula provided by the authors, where future positive returns are classified as +1 and negative or zero returns as -1.

New Section 3: Methodology Subcategories:

1. **Feature Selection (LASSO):**
   Utilizing penalized regression to identify and eliminate irrelevant features, thereby reducing model complexity and preventing overfitting.

2. **Modeling Framework:**
   Implementation of Multi-layer Perceptron (MLP) neural networks, leveraging their powerful non-linear modeling capabilities.

3. **Hyperparameter Tuning:**
   Execution of a comprehensive grid search across various hyperparameters, including network depth, learning rates, and activation functions, to optimize model performance.

4. **Indicator Optimization:**
   Application of advanced optimization techniques like genetic algorithms and particle swarm optimization to fine-tune the parameters of technical indicators, significantly improving their predictive accuracy.

5. **Validation:**
   Implementation of 5-fold cross-validation using the Jaccard distance metric, ensuring robust evaluation of the model's generalization ability.

Descriptive Statistics vs. Models

Descriptive statistics—such as Pearson correlations and standard deviations—provide foundational insights into the characteristics of a dataset. Models, by contrast, encompass predictive techniques such as LASSO and neural networks, translating the insights gained from descriptive analysis into actionable forecasts.

Optimisation of Technical‑Indicator Parameters

The authors optimise the parameters of technical indicators (e.g., moving-average periods) to align with prevailing market conditions. Such optimization keeps the parameters relevant and effective—crucial given the non-stationary nature of financial data. By continuously adjusting these parameters, the authors maintain and enhance the predictive power of their neural-network models, ultimately enabling more accurate stock-market predictions.

Q4. Feature Understanding

A feature, as defined by Sagaceta Mejia et al. (2022), is any input variable utilized by the predictive model, derived from raw market data. Specifically, features include technical indicators such as Relative Strength Index (RSI), Bollinger Bands, and Moving Average Convergence Divergence (MACD). These indicators encapsulate various dimensions of market information, including trend, momentum, volatility, and volume.

Distinguishing between features, methods, and models is crucial:

- **Feature:** Input variables extracted or computed from raw data, such as RSI or moving Averages.
- **Method:** Techniques applied to select, transform, or optimize features, exemplified by procedures like LASSO for feature selection.
- **Model:** Predictive frameworks utilizing features to produce actionable outputs, such as Multi-layer Perceptron (MLP) neural networks employed by the authors.

Categories of Features:

1. **Trend Indicators**: These include Simple Moving Average (SMA) and Exponential Moving Average (EMA), which identify the general direction of market trends.
2. **Momentum Indicators:** Indicators such as RSI and MACD, assess the speed and strength of price movements.
3. **Volatility Indicators:** Bollinger Bands and Average True Range (ATR), which measure price volatility and potential breakout points.
4. **Volume Indicators:** On-Balance Volume (OBV) and other indicators that reflect trading activity and market participation levels.

Importance of Optimization:

Optimizing features prevents overfitting by eliminating irrelevant or redundant inputs that could impair model accuracy. Given the sensitivity of neural networks to input noise, using optimized technical indicators significantly improves the predictive model's performance and generalization ability (Goodfellow et al., 2016).

Q5. Optimization Understanding

Optimization forms the backbone of robust predictive modeling, particularly in the context of financial time series where data is noisy and non-stationary. In the methodology adopted by Sagaceta Mejía et al. (2022), several core optimization techniques and evaluation metrics are utilized to ensure that the predictive models generalize well to unseen data.

Cross-Validation and K-Fold Cross-Validation:

Cross-validation is a model validation technique that assesses how the results of a

statistical analysis will generalize to an independent dataset. Specifically, k-fold cross-validation divides the original dataset into k equal-sized folds or subsets. The model is trained on k-1 of these folds and tested on the remaining fold. This process is repeated k times, with each fold used exactly once as the test set. The performance metric is then averaged across all runs, providing a more reliable estimate of the model's out-of-sample performance and minimizing the risk of overfitting (Goodfellow et al., 2016).

Jaccard Distance:

Jaccard distance is a metric used to quantify the dissimilarity between two sets. For classification problems, it is especially useful in evaluating the similarity between predicted and actual class labels. The Jaccard distance is defined as one minus the size of the intersection divided by the size of the union of two sets:

$$J = 1 - \frac{|\text{Intersection}|}{|\text{Union}|}$$

In the context of this paper, it measures how well the model's predictions align with actual market movements—a lower Jaccard distance signals better predictive accuracy.

Comparison to Other Metrics:

- **Hamming Distance:** Counts the number of mismatches between predicted and actual labels. Unlike Jaccard, Hamming distance does not account for the proportion of overlap, making it less informative for multi-class problems.
- **Cosine Similarity:** Measures the cosine of the angle between two non-zero vectors, often used for continuous or high-dimensional data, but less

interpretable for binary classification tasks.

Optimal Solution Definition:

The authors define the optimal solution as the set of technical indicator parameters and neural network hyperparameters that maximizes classification accuracy while minimizing the risk of overfitting. This dual focus on accuracy and generalization ensures that the model remains robust under varying market conditions and does not merely memorize historical data (Sagaceta Mejía et al., 2022)

PART 2:

1. Sources of Data

Behavioral data are a type of alternative data that are generated from users' behavior, patterns, and activities on digital platforms and reveal their attention and sentiments. They are short-term, real-time, and low-cost in nature.

Sources of behavioral data include:

- Social Media: These include websites like Facebook, Twitter (now X), Yelp, and Stocktwits.com. These sources are useful for monitoring trends in firm returns and stock prices, new sentiments, and instantaneous responses to occurrences owing to their real-time, large-scale, global, and flexible characteristics.
- News Articles: These encompass news articles from online news platforms such as the WTO, The New York Times, and The Wall Street Journal. News data are regarded as professional and credible, presenting an unbiased perception of market situations and business scenarios.
- Searching Volume Index (SVI): These take the volume of search terms across sites such as Google and Baidu. SVI offers a real-time, cost-effective, and worldwide quantitative assessment of public interest and sentiment on particular subject matters.

2. Types of Data

Behavioral data, one of the alternative types of data, is derived from the activities, habits, and interactions of users across different digital platforms, providing information about their attention and sentiments. This data is distinguished from others by its short-term, real-time, and low-cost nature.

The primary categories of behavioral data are:

- **Social Media Data:** This is data gathered through social networking platforms such as Facebook, Twitter, and Yelp. It is valuable for monitoring stock price trends and firm returns, discovering new sentiments, and learning

real-time responses to events. It is large-scale, independent, efficient, global, and flexible in nature. Some specific sub-types of social media data include:

  - **Connection-based networks:** Sites such as Facebook and Twitter, where users freely provide opinions and form online connections, offering a balanced view of the public's opinion. These are primarily used for stock price prediction based on sentiment analysis of firms and gauging company returns from social media updates.
  - **Interest-based networks:** Sites like Stocktwits.com or Scutify.com, which are dedicated to a specific subject like finance, attract users who are enthusiastic and experienced in those areas. Investors, expert analysts, and businesspeopletend to use these networks, and their discussions provide useful information on market direction and stock performance.
  - **Review-based networks:** Web platforms that allow users to analyze and rate products, businesses, or services, often anonymously, to ensure credibility. They are effective for forecasting future product sales, company performance, and specific stock prices.

- **News Stories:** These are facts from reports or stories released by third-party professionals, i.e., online news portals. The high credibility of the publishers makes news facts objective and trustworthy to investors and businesses. News is released regularly and broadly, affecting stock prices by changing investors' attitudes and mitigating management risks by tracking companies.
- **Searching Volume Index (SVI):** It provides quantitative measurement of public sentiment and interest toward particular subjects. Google and Baidu are providers of SVI data. SVI is timely, cost-effective, and worldwide, enabling investors to understand existing market trends and predict future market dynamics.

3. Quality of data

This user guide subsection addresses the quality of behavioral data and highlights the typical problems and mitigation strategies, as presented in the research paper provided.

Quality of Behavioral Data

Although behavioral data presents enormous benefits in its timeliness and capacity to capture real-time emotions, the quality of this data is of paramount importance to sound financial and business analysis. Various issues inherent to the nature of such data may affect its trustworthiness and require utmost care to treat.

Common Quality Issues

Noise and Irreducible Errors: Behavioral data tends to be raw and unstructured, containing by necessity "noisy" information. Such noise creates statistical errors, in the form of Type I errors (false positives, wrongly rejecting a true null hypothesis) and Type II errors (false negatives, wrongly accepting a false null hypothesis).
False, Exaggerated, Commercial, and Unofficial Information: Social media, being a primary behavioral data source, is susceptible to false, exaggerated, commercial, or unofficial information. For instance, it has been observed through research that a significant percentage of tweets (approximately 1-14%) are from "bot" accounts, which can spread misinformation. The rise of new NLP technologies such as ChatGPT and LLaMA only adds to the challenge of separating real information from created or manipulated information.

- **Intrinsic Bias:** Behavior data typically comes with some level of bias, especially since it is affected or created by individual perspectives.
- **Exaggeration and Social Conformity:** People tend to exaggerate their views or present favorable content selectively because they are worried about adverse consequences (e.g., losing their job, loss of opportunities) or because they want to be socially accepted. This results in biased sentiment or opinion data.
- **Partial or One-Sided Information:** Posts and comments may be driven by incomplete or skewed information, causing imbalanced interpretations.
- **Selection Bias:** No one source or form of behavioral data can completely account for the entire population or all applicable entities (persons, firms). When big data sets are analyzed based on just a sample of the data, conclusions made might not reflect the larger population, causing erroneous or incomplete interpretations.

- **Incompleteness and Timeliness:** Behavioral data tend to be timely, but individual data points may be incomplete or may arrive with different lags, impacting the completeness of the analysis. Representativeness may also be a problem if the data only reflects the behavior of a niche segment and not a large market segment.

Methods for Evaluating and Reducing Quality Issues

Resolving these issues of quality is important to derive actionable insights from behavioral data. The paper proposes some methods to evaluate and prevent these problems:

- **Data Cleaning and Preprocessing:** Central to enhancing data quality is effective data cleaning. This entails the elimination of irrelevant content, the treatment of missing values, normalization of formats, and normalization of text (e.g., lowercasing, punctuation removal, tokenization) to facilitate analysis. For text data, this also comprises methods used to detect and exclude spam, off-topic posts, or machine-generated content.
- **Statistical Methods for Bias Mitigation:**
- **Robustness Checks:** Using alternative, possibly more conventional, data sources to cross-check conclusions obtained from behavioral data can be used to confirm conclusions and detect possible biases.
- **Stratified Random Sampling**: In the case of large datasets, using stratified random sampling makes certain that chosen subsets are representative of the entire population, thus avoiding selection bias.
- **Anonymous Users:** Making it easier to encourage or provide data collection anonymity can minimize the "herding effect" and prevent the amplification or social conformity of opinions.
- **Technological Approaches to Bias Detection**: Artificially intelligent algorithms utilizing advanced Natural Language Processing (NLP) methods may be used to identify and measure bias in textual behavioral data. Studies of more advanced approaches to detection of bias in a diversity of data formats (text, image, audio, video) continue to be ongoing and indispensable due to the dynamic characteristics of data creation.
- **Bot Filtering and Detection:** Due to the high occurrence of bot accounts, particularly on social media, installing advanced bot detection filters is

crucial to eliminate non-human generated content that can skew sentiment and volume analysis.

- **Contextual Knowledge**: Analysts should have rich understandings of financial markets and the particular businesses under analysis in order to be able to properly read the value of alternative data and distinguish between valuable signals and simple noise. Integrating multiple forms of alternative data can also increase investigation scopes and present more complete views, to cross-validate data where single forms of data may have their limitations (e.g., combining satellite imagery with social media postings to gauge disasters).

4. Ethical Problems with Behavioral Data

The application of behavioral data, especially from social media, in finance is fraught with a number of ethical problems, as much as it has analysis potential. Resolving these problems is essential for appropriate data use.

Major Ethical Problems

- **Privacy Breaches:** A key issue is personal privacy. Despite anonymization, there is still a danger of re-identification of individuals by cross-referencing different pieces of information. Consent for trading on public social media information for financial analysis is usually tacit and not properly informed, with most people unlikely to knowingly provide sensitive personal information for that purpose. Data leakage remains a major risk.
- **Data Ownership and Control:** Ownership of social media user-generated data is complicated. Users tend to have little direct control over how their data is being aggregated, disseminated, or sold to third-party vendors for financial uses, although the data is "publicly generated".
- **Potential for Misuse:** Market Manipulation: Social media's ubiquity and immediacy can enable market manipulation, in which fake accounts or coordinated manipulation disseminate false news to manipulate share prices, creating market instability. Such activities can be detected using behavioral data as well.

- **Discrimination:** Abuse of behavioral data may result in discriminatory behavior in credit assessment or investment choices, premised on correlations with sensitive characteristics.
- **Surveillance:** Mass collection of data, even for financial purposes, is a concern regarding digital surveillance and individual freedom.
- **Bias Reinforcement:** Biases inherent in behavioral data (e.g., selection bias, overamplified content, echo chambers) have the potential to be enhanced if not adequately mitigated and may result in discriminatory or misleading financial models and results.
- **Regulatory Compliance:** It is tricky to comply with changing data privacy regulations such as GDPR (mandating explicit opt-in and offering broad individual rights) and CCPA (offering know, delete, and opt-out of data sale rights for California residents). Regulatory compliance like that of HIPAA might even be applicable if sensitive financial information is involved.

Ethical Mitigation Strategies

- **Privacy-Preserving Technologies:** Utilizing technologies such as "privacy computing" is essential. This includes encryption and other means to allow data analysis without divulging real sensitive information, hence decreasing leakage risk and spurring data sharing.
- **Transparency and Informed Consent:** Be transparent about data collection and usage practices with users. It is essential to obtain explicit and informed consent, particularly for sensitive data or new financial uses.
- **Strong Anonymization and Aggregation:** Adopt sophisticated methods to anonymize data and aggregate data to reduce re-identification threats. A shift towards analyzing aggregate trends instead of individual-level data also provides added privacy.
- **Active Bias Management:** Proactively utilize statistical approaches (e.g., robustness checks, stratified sampling, promoting anonymity) and technical approaches (e.g., NLP-based bias detection) to detect and rectify biases in behavioral data sets. This maintains fairness and accuracy in models.
- **Strong Data Governance and Security:** Have well-defined data ownership policies in place and apply strict security protocols to avoid unauthorized access and breaches.

- **Adherence to Regulations:** Be in constant adherence with data protection laws like GDPR and CCPA, grasping their provisions regarding data minimization, purpose limitation, and retention.
- **Ethical Review:** Have internal ethical review procedures to review new uses of behavior data, especially within finance, to actively spot and address possible ethical issues.

Short Literature Search Linking to Related Research

This section summarizes academic literature on social media (behavioral) data in finance, connecting it to the review by Sun et al. (https://www.google.com/search?q=2024).
Sun et al. cite_start extensively review social media's role, highlighting its real-time nature and utility for tracking trends and sentiments, categorizing it into connection-based, interest-based, and review-based networks for applications like stock price forecasting and risk assessment.

Key Related Research:

- **Social Media Sentiment & Stock Prices:**
  - Bartov et al. cite_start suggest Twitter opinions correlate with company earnings and stock price reactions.
  - Kraaijeveld and De Smedt (2020) used Twitter sentiment to predict cryptocurrency returns.
  - Siganos et al. cite_start found Facebook sentiment divergence impacts stock volatility and trading volume.
  - Chen et al. cite_start showed comments on investor platforms like Seeking Alpha influence stock prices.
  - Huang (https://www.google.com/search?q=2018) indicated Amazon reviews predict stock returns and earnings.
- **Social Media & Corporate Performance/Behavior:**
  - Ge et al. cite_start observed that influential tweets could boost stock prices and investor attention.

- Wang and Chen (2020) linked CEO personalities derived from social media posts to cost efficiency and employee productivity.
- Grover et al. cite_start used Twitter to assess corporate social responsibility's impact on reputation.
- Stamolampros et al. cite_start connected Glassdoor employee reviews to job satisfaction and return on assets.
- Quinton and Wilson (2016) found LinkedIn participation improved business relationships and performance.

- **Challenges & Future Trends:**
  - Sun et al. cite_start address challenges like noise, various biases (selection, exaggeration, bots), and interpretation difficulties. They foresee future needs for analyzing image/audio/video data and the role of privacy computing.

References:

Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using

attention mechanism. IEEE Access, 7, 11373–11386.

BlackRock. (2024). iShares Core S&P 500 ETF. Retrieved from

https://www.ishares.com/us/products/239726/ishares-core-sp-500-etf

Chong, T. T., & Ng, W. K. (2008). Technical analysis and the London stock exchange: Testing

the MACD and RSI rules using the FT30. Applied Economics Letters, 15(14), 1111–1114.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Han, Y., Yang, K., & Zhou, G. (2021). Technical trading indicators and the stock market:

Evidence from a computational perspective. Journal of Financial Markets, 54, 100601.

Sagaceta Mejia, J., Vallejo Huanga, D., & Paredes Valverde, M. A. (2022). An intelligent

approach for predicting stock market movements in emerging markets using optimized technical

indicators and neural networks. Economics, 16(1), 122–142. https://doi.org/10.1515/econ-2022-

0073

Yahoo Finance. (2024). IVV historical data. Retrieved from

https://finance.yahoo.com/quote/IVV/history

Zhang, Y., Wang, S., & Wang, L. (2021). Financial time series forecasting with deep learning: A

systematic literature review. Applied Sciences, 11(9), 3856.