EFFICIENT VISUALISATION AND PREDICTIONS FOR SPACE WEATHER APPLICATIONS

1. INTRODUCTION

1.1 Background

Space weather refers to the planet's dynamic conditions, which are driven by solar activities in the planet's magnetosphere, ionosphere, and thermosphere (Narechania et al.,2020. Licata, Tobiska, & Mehta, 2020). However, science has evolved to develop patterns and structures to understand the future possibility of weather conditions across the globe via forecast, especially with the growing threats posed to the Earth's surface. Weather forecasting and exploration across other planets have been explored as they play a crucial role in various facets of human daily life. Granda and Dharavath (2021) explained that weather forecasting is employed to anticipate the future atmospheric state of a designated geographical area.

Hassler et al. (2022) explained that although solar activity has been declining as the Sun approaches solar minimum, a series of large solar storms occurred in September 2017 that impacted both Earth and Mars

The effects and impact of space weather on other planets are becoming increasingly important as space research and human exploration expand into the solar system. One of the planets being explored is Mars. Unlike Earth, the surface of Mars is much more exposed to space radiation and the effects of space weather. This is true because Mars lacks a global magnetic field or magnetosphere to deflect high energy charged particles, and the Martian atmosphere is very thin (roughly two orders of magnitude smaller column density than Earth), providing significantly less effective shielding. As a result, exposure to the radiation environment on the surface of Mars remains a significant concern and health risk for future human explorers. The radiation environment can also be dominated, on short time scales (usually hours to days), by SEPs generated at the Sun and accelerated by solar flares or associated shocks. The possibility of exploring predictive studies based on the data available has been a viable way to understand the planet and the potential of having human life on the planet.

Currently, the field of meteorology that entails space science and weather application relies on physics-based numerical weather prediction techniques to forecast weather conditions. (Punjabi, 2021). However, implementing these methods requires substantial computational resources, which might sometimes be challenging or fraught with human errors. Accurate prediction and efficient visualisation of space weather phenomena are crucial for various applications, including satellite operations, communication systems, and power grids. Gong et al. (2022) explained that the accurate prediction of meteorological conditions is important across various domains in contemporary society. They possess significant importance in the realms of economy and industry, such as agriculture and utilising renewable resources. The electric power industry is crucial for both the provision of electricity and the mitigation of natural hazards. According to the study, Numerical weather prediction (NWP) models have been operationalised at meteorological centres worldwide since the early 1960s. Currently, these models can simulate the dynamics of the global atmosphere at a resolution as far as one kilometre (Gong et al., 2022). Although the accuracy of their predictions has achieved an extraordinary level, the computational resources necessary for these predictions are substantial.

The beginning of the 21st century, with the emergence of massive data, effective supercomputers with Graphics Processing Units (GPU), and scholarly interest in emerging new methodologies, turned out to be a pivotal time in the history of machine learning (Fradkov, 2020). With tremendous data volume and computer power growth, recent years have been regarded as the golden era for artificial intelligence and machine learning. This is despite the fact that many approaches have been known since the 1960s and have been investigated in detail in many papers since that time (Bochenek and Ustrnul, 2022).

In recent years, there has been a notable proliferation of deep-learning models that have achieved considerable success in various tasks on weather prediction. The advent of the big data era presents opportunities for substantial enhancements in the accuracy of weather condition predictions. Machine learning algorithms have emerged as powerful tools for space weather prediction because they can analyse complex datasets and capture intricate patterns (Bochenek & Ustrnul, 2022). Techniques like deep learning and U-Net architecture have shown promise in improving spatiotemporal weather forecasting by efficiently capturing spatial dependencies (Punjabi & Izquierdo-Ayala, 2021)

Mahesh (2019) stated that some themed papers contain detailed assessments of machine learning algorithms. Machine learning algorithms are the most important subgroup of artificial intelligence technologies in atmospheric research. Bochenek and Ustrnul (2022) explained that two different approaches exist when predicting using machine learning, which is said to include the supervised and unsupervised machine learning approaches. Supervised learning, which was determined to be the most dominant group in the most recent papers on the field of atmospheric science, was deemed to be the most attractive group of approaches for atmospheric scientists (Lugaz et al., 2022). If some labelled data are accessible, one can utilise the data as a training dataset to construct a function that maps the inputs provided to the desired outputs. That function can be used in a new dataset, dubbed testing, to evaluate the model. If the results are satisfactory, they can be used in the classification or regression of any required application. The second type of machine learning is unsupervised learning, and it is different from supervised learning in that it does not use labelled data to train algorithms. Instead, unsupervised learning requires algorithms to determine alternative ways to lower the dimensions of a dataset to conduct further research.

In contemporary times, there has been significant advancement in technology, particularly in the domain of Machine Learning (ML), which has proven instrumental in alleviating human labour. Within artificial intelligence, machine learning (ML) amalgamates principles from statistics and computer science to construct algorithms that exhibit enhanced efficiency through exposure to pertinent data instead of relying solely on explicit instructions. In addition to speech recognition, image detection, and text localization, machine learning (ML) involves investigating computational algorithms autonomously improved through experiential learning. The classification of artificial intelligence as a subset is widely acknowledged in academic literature (Abdulqader et al., 2020). To generate predictions or make decisions, machine learning algorithms have the ability to create a model population using a sample dataset, commonly referred to as 'training data' (Libbercht & Noble, 2015). Machine learning (ML) algorithms are widely employed in various domains, such as email filtering and computer vision, to address challenges associated with developing conventional algorithms for implementing required functions (Carleo et al.,2019). Machine learning (ML) encompasses a wide range of applications, one of the most notable being predictive data mining. Two primary mechanisms in ML classification fulfilment are model development and model evaluation (Hillel et al.,2020).

Every dataset utilised by machine learning algorithms describes each instance using the exact attributes. The attributes may fall into three categories: continuous, categorical, or binary (Zeebaree et al.,2019). Supervised learning is characterised by recognising cases through recognising recognised labels, which serve as correct outputs (Sharma & Kumar, 2017). Supervised learning is a fundamental machine learning task involving inferring a function from a pre-classified training data set. Additionally, the provided algorithm examines the testing data and generates a derived task that can be employed to map new examples (Sharma & Kumar, 2017). However, it is essential to note that each data input object is pre-assigned with a class label. The main objective of supervised algorithms is to acquire a model that can accurately assign labels to provided data and generalise well to new, unseen data. Classification aims to predict the target class with the utmost precision accurately. The classification algorithm determines the relationship between input and output attributes to create a model through training (Pahwa & Agarwal, 2019).

Lugaz et al. (2022) explained that although machine-learning approaches in space weather have been around since the field's early years, there has been a noticeable increase in published studies utilising machine-learning techniques in the last four years. This development can be attributed to the fact that machine-learning techniques are becoming increasingly popular. This advent of studies and enquiries into using machine learning for prediction has led to a debate about ML-based forecasts of space weather occurrences and the roles it is expected to play in space weather research.

One of the primary difficulties encountered in prediction tasks involving space weather data is determining the appropriate machine-learning technique for conducting these predictions. Most machine learning techniques necessitate a substantial quantity of data, which can pose challenges in data acquisition and generation, as selecting or creating an appropriate database may not always be straightforward. (Obando et al., 2021).

It is essential to point out that Camporeale (2019), in a review of Grand obstacles, explored the promise and technical barriers of applying machine learning to research on space weather. It was observed that having a perfect understanding of probabilistic forecasting and analysing uncertainties is essential for its applications. Morley (2020) analysed that it is vital to understand ML techniques in space weather forecasts with the goal of better understanding the factors that influence space weather. One of the most significant challenges in evaluating the progression or improvement of a new machine learning approach is the absence of a suitable comparison with published results obtained using existing methods. It is necessary to make available an easy way of intercomparison between the various research conducted in space weather to allow the correct application and continuous development of machine learning prediction.

Another critical component of understanding weather forecasts is visualisation techniques. Visualising space weather data is crucial in understanding and interpreting complex patterns and trends. Integrating machine learning with data visualisation techniques can facilitate exploring and analysing large-scale space weather datasets (Granda & Dharavath, 2021).

Hepworth et al. (2022) explained that developing visualisation abilities is paramount for students studying space weather forecasts. This skill is essential for constructing precise mental representations of fundamental concepts, evaluating the outcomes of various experiments, and effectively conveying findings to wider audiences. With the increasing volume of data and the growing reliance on computing power, this visualisation has become even more essential in various aspects of geoscience research and practice. According to Malakar et al. (2020), critical weather applications necessitate online visualisation performed concurrently with the simulations so that scientists may give decision-makers real-time assistance. To support the understanding of how visualisation techniques are built and applied to machine learning pipelines, Yuan et al. (2021) noted that exploring visualisation techniques is now necessary due to the rapid growth of visual analytics approaches for machine learning. There have been several early attempts to compile the developments in this field from various angles, but this has become more important in present times.

From the preceding, Deep learning (DL) has been successfully used in computer vision applications recently, including anomaly detection, human action prediction, and self-driving automobiles. These demonstrate that machine learning can identify intricate patterns and unearth incredibly nonlinear relationships using data-driven methods. As a result, there is optimism that deep learning can be applied to Earth system science and weather forecasting, which both deal with several complex, multi-scale, and non-linear coupled processes (Schultz et al., 2021). In the context of weather and climate forecasting, emulation of physical parametrisation (e.g., Han et al., 2020), and detection of extreme weather events in climate datasets (e.g., Racah et al., 2016), the weather and climate communities are starting to research the use of these advanced machine learning (ML) methods, more so, as there is an extension with the possibility of exploring new frontiers with Planet space. Several possible DL applications in weather forecasting are covered in Schultz et al. (2021). The NWP workflow entails the preprocessing of observational data, assimilation of these data into the simulated actual atmospheric state, forecasting using a numerical model, and post-processing of the raw model outputs, which can incorporate machine learning approaches (Schultz et al., 2021).

However, despite advancements in machine learning and visualisation, there is still a need to enhance the efficiency and accuracy of space weather predictions. Additionally, there is a lack of research focusing on the efficient visualisation and prediction of space weather phenomena, especially in analyses outside Earth space (Meza-Obando, Salas-Matamorros, Mariño, & Lopez, 2021). This study aims to develop an efficient visualisation and prediction framework for space weather applications by addressing these gaps and leveraging machine learning techniques on Mars. The proposed framework will improve space weather forecasting accuracy, enabling better decision-making and mitigation strategies for various sectors that rely on space-based technologies (Narechania et al., 2020).

1.2. Research Aims and Objectives

This project aims to provide a data-driven solution to the problem of predicting Mars weather, including the temperature and amount of ultraviolet radiation.

Specifically, the study aims to:

- Perform feature engineering on the datasets.
- Train a Deep learning model for the temperature data prediction and a Decision Tree Classifier for the UV Radiation Prediction.
- Analyse the model's performance and, where applicable, finetune the model to perform better.
- Provide solutions to predict the next temperature values given the current or previous values and to predict the UV Radiation, given the temperature.

2. LITERATURE REVIEW

2.1 Existing methods and techniques for visualising and predicting space weather

Space weather is the ever-changing condition of space, and it is heavily influenced by solar activities and interactions with planets' magnetic fields. These interactions cause a range of challenges for modern technology in space studies. It also makes the precise visualisation and prediction of space weather important because it plays a vital role in protecting communication and navigation systems, satellites, and power grids. This review of current methodologies and approaches for visualising and forecasting space weather is focused on machine learning, data-driven models, and visualisation technologies.

Over time, machine learning methodologies have attracted academic interest as a viable method of predicting space weather. In their study, Reiss et al. (2020) thoroughly examined different machine-learning techniques focusing on artificial neural networks and ensemble methods for predicting space weather. The study by Srivastava et al. (2018) also focused on using machine learning techniques to forecast geomagnetic storms based on solar wind data, leading to enhanced precision and advanced warning capabilities in their predictions. It is important to note that the most common studies in space weather have been related to Earth space, but recent development has shifted attention to studying other planets.

One potent methodology for visualising space weather is the data-driven solar wind model. Kilpua et al. (2017) presented a novel 2.5D data-driven solar wind model that is a valuable tool for investigating space interactions. These models contribute to our understanding of solar-wind interactions and their consequences for space weather, allowing researchers to mitigate weather hazards. The model is similar to the framework proposed by Orange et al. (2018), which employed deep learning algorithms to generate precise solar image predictions that showed solar flares and coronal mass ejections for monitoring factors influencing space weather.

The Nychka et al. (2019) study investigated virtual reality (VR) for visualising weather data. The study emphasised the capacity of VR-enabled methods to improve the processing of data exploration and analysis. Using virtual reality technologies, researchers engage in an immersive research experience within 3-dimensional visualisations of space. This allows for dynamic interactions and helps make informed conclusions about space weather events. Gallagher et al. (2016) used Python-based models to explore space weather data interactively. The flexible and robust libraries of the Python programming language allow for an intuitive interpretation of delicate space weather datasets, thereby improving data-driven prediction and analysis of space weather.

To improve space weather forecasting, Pulkkinen et al. (2017) created a novel analogue ensemble method that used past data to make predictions about future weather events. This method used historical space weather occurrences to identify comparable circumstances and offer valuable insights for predicting space weather, especially when machine learning methods are constrained in accessing data for predictions.

The research by Jensen et al. (2019) explored the concept of ensemble forecasting, which integrated multiple predictions to generate dependable forecasts. The approach sought to reduce uncertainties in space weather predictions by using different methods to improve their precision and resilience.

Further studies on the existing methods and techniques for visualising and predicting space weather continue to be an area of academic interest because they present novel prospects for anticipating and mitigating factors influencing space weather occurrences. This becomes particularly interesting for this study.

2.2 Space weather prediction

The study of space weather prediction focuses on understanding and predicting the dynamic interactions between the Sun and the Earth's magnetosphere, ionosphere, and atmosphere, and in recent times, the other planets and the Sun.

In their study, Bobra and Couvidat (2015) presented a machine-learning algorithm that used the Solar Dynamics Observatory/Helioseismic and Magnetic Imager data to predict solar flares. Solar flares are sudden releases of energy from the sun that may affect technologies in space. Therefore, space agencies and satellite operators must increase their ability to predict the occurrence of solar flares and proactively implement preventive strategies. Riley (2018) presented insights into the statistical analysis of infrequent but potentially critical space weather incidents to comprehensively understand the probability of such occurrences. This is important for effectively designing proactive measures to safeguard essential infrastructure and space technologies. They also underscored the need for collaboration among governmental bodies, industries, and researchers to develop a robust system that can withstand the adverse effects of space weather events like geomagnetic storms that can generate geomagnetically induced currents within power grids. If GICs occur, they can trigger widespread power outages and subsequently disrupt human activities on a large scale (Boteler & Pirjola, 2017).

Similarly, identifying vulnerabilities in technological systems and developing a resilient prediction model is necessary for mitigating potential damages to space technology. Siscoe and Crooker (2017) posited that with precise predictions, satellite operators could be empowered to make strategic manoeuvres to minimise exposure to charged particles and electromagnetic radiation. This will help prevent potential disruptions in global communication and navigation services. Power grid operators can also use real-time forecasts to proactively implement measures like modifying power distribution plans to safeguard transformers and other intricate elements against the destructive effects of geomagnetically induced currents (Pulkkinen et al., 2017).

For Schrijver and Kauristie (2015), recognising the relationship between the Sun and the Earth is essential for space weather forecasting because solar activities substantially influence the Earth's magnetosphere. Their study underscored the importance of understanding solar activities for enhancing predictive models of space weather events. By integrating solar inputs into space weather models, researchers may be able to predict the timing and magnitude of solar storms to quickly notify spacecraft and astronauts, thus ensuring their safety during space expeditions. There are potentialities of carrying out the same on other planets.

Leka and Barnes (2017) recommended the McIntosh Archive as a source of valuable information for analysing global solar magnetic field maps. Global solar magnetic field maps assist in a deeper understanding of solar magnetic patterns and provide data for improved space weather prediction. Using these advanced observational tools and datasets facilitates the analysis of space weather events. It enhances the understanding of the fundamental mechanisms involved in refining the precision of predictive models.

To adequately protect space technological infrastructure and improve readiness for space weather events, it is vital to prioritise ongoing research efforts to establish robust prediction models. The effect of space weather advancements on everyday life also makes the development of accurate space weather prediction models essential for protecting society.

2.3 Available data on space weather

Examining the worldwide dissemination of Coronal Mass Ejections (CMEs) and CME-driven shocks, a study by Bisi et al. (2018) emphasised the significance of using multiple data sources in understanding the intricate features of space weather. The study used data from the STEREO and Wind missions and various observatories to successfully monitor the spatial and temporal progression of coronal mass ejections (CMEs) and their consequential effects on space weather. Webb and Howard (2012) also analysed the coronal mass ejections and the observatory models used to study them. They emphasised the importance of accurate and ongoing data acquisition and monitoring for anticipating and mitigating the effects of space weather events. Their study concluded that researchers should combine space-based observatories and ground-based instruments to effectively survey solar activities and their potential effects on space weather.

Satellite mission expeditions have also played a significant role in acquiring space weather data. Bothmer and Daglis (2017) analysed the physics and impact of space weather, focusing on the importance of satellite observation in investigating solar activities. The study found critical insights into the sun's behaviour, like solar flares and coronal mass ejections, which may affect the magnetosphere and ionosphere of Earth. This is important for minimising the potential harm to terrestrial power grids.

Similarly, Gopalswamy (2006) also researched some unresolved questions in the space weather investigation. While emphasising the necessity of extensive datasets for tackling lingering inquiries, they concluded that it is important for researchers to conduct extensive observations and monitor space weather to establish patterns and trends. The importance of continuously obtaining data is to improve the ability to predict space weather events and identify early warning signs.

The importance of data-driven models for spatial understanding has been underscored by Schrijver and Siscoe (2010). In their study, they explored the field of heliophysics. They focused on the analysis of predictive models and data techniques that have been used to anticipate weather events and their consequences. They encouraged the use of sophisticated numerical stimulators to generate weather forecasts accurately. These models have played a crucial role in predicting solar storms, solar radiation hazards, and geomagnetic disturbances caused by space weather events.

Similarly, Coates and Hapgood (2012) studied the role of physics as an underlying cause of space weather phenomena. They emphasised the significance of acquiring comprehensive data to comprehend the intricate nature of space weather. They encourage interdisciplinary research in solar physics, magnetospheric and ionospheric physics, and atmospheric science for a thorough understanding of space weather.

Understanding the interdependence of space weather and Earth's atmospheric and magnetic conditions, Reiff and Scaefer (2010) presented the necessity of using a comprehensive approach to studying space weather. They underscore the importance of timely data exchange and collaborative efforts among international space organisations to achieve a complete understanding of space weather events on a global level. Establishing interdisciplinary research and collaboration between solar scientists and engineers will potentially enhance space weather forecasting and improve the overall resilience of technology-dependent societies. Analysing existing space weather data will further establish a solid basis for future scholarly investigations, policy formulation, and an enhanced preparation strategy for effectively addressing the complex nature of space weather events.

According to Priyadarshini and Puri (2021), the colonisation of Mars has been spurred on by human curiosity and the potential for habitation on other planets to lessen the chance of human extinction. Many public space agencies support planet-wide exploration, colonisation, and human missions. Polar ice sheets give scientists great hope that they may make Mars habitable, although many conditions prevent colonisation, such as toxic soil, low gravity, radiation exposures, etc. There seems to be constant discussion over how to make the planet suitable for humans when considering colonising it. These considerations include the atmosphere, soil, water content, and other elements. Hence, they stated the need to investigate various factors that may support Mars' colonisation to enhance or undermine the claim. One such issue worth exploring is the weather.

The study by Garcia et al. (2022) explained that data-driven methods were used in making forecasts using artificial intelligence to deterministic methods employing sophisticated fluid dynamics models. The first is primarily concerned with developing General Circulation Models, whilst the latter is beginning to take place in many contexts for Earth's meteorology and astrophysics. The study used environmental data from the Vikings and Mars Science Laboratory missions to train an artificial neural network to forecast the weather on Mars. The methodology used in this work is a data-driven approach, employing computer science expertise that has long been used on Earth but has not previously been utilised on Mars.

2.4 Previous Studies on Space Weather

In the research by Priyadarshini and Puri (2021), some artificial intelligence methods for studying meteorological data from Mars were discussed. We use machine learning models like CNN, CNN-LSTM, GRU, stacked LSTM, layered LSTM, and Long Short Term Memory (LSTM) to analyse the meteorological data for the red planet. The models were validated using statistical variables, including MAE, MSE, RMSE, and R-squared coefficient. According to the findings, the LSTM model performs better than all baseline models, with an R-squared value of 0.8640 and an MAE value of 0.1257.

The study by Garcia et al. (2022) built an artificial neural network that uses data from the previous day as input to forecast the weather for the following day. The study demonstrated that temperature and pressure are two of the most crucial factors and that ANN can predict daily changes in the chosen variables with an accuracy of 0.5 to 1%.

According to Lugaz et al. (2022), over the past few years, there has been a noticeable surge of manuscripts in Space Weather that use machine learning approaches. The study highlights those manuscripts focusing solely on a forecasting technique (as opposed to understanding and forecasting a phenomenon) that must substantially improve the current state-of-the-art methods. We present this comparison by discussing which manuscripts fall within the journal's scope. The study showed that data preparation details must be included in every paper, including how the data were divided into training, validation, and testing sets. At the time of submission, the software and algorithms used to create the machine-learning technique should be stored in a repository. It is necessary to compare published results obtained using alternative approaches and go over the projected results' uncertainty. If they demonstrate a significant advance over existing state-of-the-art forecasting models, manuscripts describing a new forecasting scheme based on an existing ML technique are in scope. The new model's results must be contrasted with those from contemporary models in the manuscript. Unless there are no more suitable models, the comparison cannot be restricted to the most basic models (climatology, persistence, etc.).

One significant barrier to evaluating the development or improvements of a new ML approach has been the absence of adequate comparison with published results utilising previous methods. Simple intercomparison between various research must be made possible to assist the practical application and continued advancement of ML in space weather.

Kashinath K et al. (2021) showed that Machine learning (ML) provides unique and powerful techniques for reliably and efficiently recognising complex patterns, simulating nonlinear dynamics, and predicting the spatiotemporal evolution of weather and climate phenomena. These are just some of the many applications that ML may be used for. On the other hand, standard machine learning models do not always adhere to the fundamental rules governing physical systems' operation, nor do they generalise particularly well to situations for which they have not been taught. We explore systematic ways of combining physics and domain information into machine learning models and then distil these ways into major categories. Ten different case studies demonstrated how these methods have been successfully utilised for simulating, downscaling, and forecasting weather and climate systems. These investigations have resulted in several successful outcomes, including increased physical consistency, decreased total training time, improved data efficiency, and enhanced generalisation. Finally, a summary of the key takeaways and outline of the scientific, diagnostic, computational, and resource obstacles that must be overcome to construct physics-informed machine learning models that are robust and trustworthy for weather and climate processes was done. This article is one of several included in the issue with the theme "Machine learning for weather and climate modelling."

Camporaele et al. (2018) explained that their work brought together experts from the fields of computer science, mathematics, statistics, and space weather to discuss the use of cutting-edge methods like deep learning, information theory, and machine learning better to understand the Sun-Earth system and forecast space weather. The community agrees that creating interdisciplinary collaborations is the most effective way to fully use the potential of these cutting-edge techniques in resolving Space Weather-related issues, even though individual efforts have been made in this direction. The use of pattern recognition, deep learning, and feature selection techniques in the absence of more physics-based models for forecasting specific SW conditions (such as the occurrence and characteristics of solar flares, Coronal Mass

Ejections, energetic particles, and solar wind conditions at L1), as well as data mining techniques, are among the key findings.

Bochenek and Ustrunul (2022) worked on Machine learning in Weather Prediction and Climate Analyses. The study searched the Google Scholar search engine for 500 scholarly articles that were the most relevant to machine learning methods used in climate and numerical weather prediction since they were published in 2018. In the research on numerical weather prediction, photovoltaic and wind energy, atmospheric physics, and processes were investigated; in climate research, parametrisations, severe events, and climate change were investigated. The abstracts' most common subjects of interest were selected, and some were studied extensively. It was also possible, using the database that was created, to extract the most frequently researched meteorological fields (wind, precipitation, temperature, pressure, and radiation), methods (Deep Learning, Random Forest, Artificial Neural Networks, Support Vector Machine, and XGBoost), and countries (China, the United States of America, Australia, and India) in these areas. The authors perform critical evaluations of the existing literature to predict the future research direction of these fields, with the primary conclusion being that machine learning methods will be an essential component of future weather forecasting.

Al-saad et al. (2022) explained that one of the most popular research fields in planetary science is surface temperature forecasting over Mars since getting data on a planet's meteorological data without directly arriving on its surface is crucial. This work presents an in-depth exploratory analysis and temperature prediction for Mars as a time series problem. Based on the temperatures observed between 2012 and 2017, a Residual CNN-LSTM is developed to forecast the temperature for 2018. The experiment's findings are intended to examine four possible outcomes: forecasting the maximum temperature using terrestrial dates, predicting the minimum temperature using terrestrial dates, and the remaining two aspects predicting the same using Solar Days (SOL). When the proposed model is contrasted with other widely adopted methods, its superiority is demonstrated after evaluating the Root Mean Squared Error (RMSE).

2.5. Recurrent Neural Network

A recurrent neural network (RNN) is a type of network architecture incorporating cyclic connections within its network structure (Jurasky & Martin, 2023). This characteristic enables the network units to depend on their previous outputs, either directly or indirectly, as inputs. Although possessing significant computational capabilities, these networks present challenges in interpretability and training. Nevertheless, specific architectures have demonstrated remarkable efficacy in language processing among the broader category of recurrent networks. This section examines a category of recurrent networks known as Elman Networks (Elman, 1990) or simple recurrent networks. These networks possess inherent utility and serve as the foundation for more intricate methodologies such as the Long Short-Term Memory (LSTM) networks expounded upon subsequently in this chapter. The term "RNN" denotes the simpler and more restricted networks (although it is worth noting that the term "RNN" is frequently used to encompass any network with recurrent characteristics, including LSTMs).

Similar to conventional feedforward networks, the input vector xt, which represents the current input, undergoes a multiplication operation with a weight matrix. Subsequently, the resulting product is subjected to a non-linear activation function to calculate the values for a layer of hidden units. The concealed layer is later utilised to compute a corresponding output, yt. In contrast to our previous methodology, which relied on a window-based approach, the current procedure involves processing sequences by sequentially presenting individual items to the network. In this context, subscripts will denote time, whereby xt will denote the input vector x at a given time t. The primary distinction between a feedforward network and the network depicted in the figure is the presence of a recurrent link, denoted by the dashed line. The provided hyperlink enhances the input to the computation at the hidden layer by incorporating the value of the hidden layer from the previous time step. The concealed layer from the preceding temporal step offers a type of memory or context, that encodes previous processing and influences the subsequent decisions to be made at subsequent temporal instances. This approach is characterised by its lack of a predetermined limit on the length of the prior context. The previously hidden layer can encapsulate information that extends to the inception of the sequence. Including this temporal dimension renders recurrent neural networks (RNNs) seemingly more intricate than non-recurrent architectures. However, they exhibit minimal dissimilarities. The primary alteration resides in the novel array of weights, denoted as U, which establishes connections between the hidden layer at the preceding time step and the current hidden layer. The weights play a crucial role in influencing the network's utilisation of the previous context when computing the output for the present input. Like the other weights in the network, these connections are trained through backpropagation.

2.5.1. Long Term Short Memory

This neural architecture has gained significant traction and widespread usage, leading to its integration within different industries. Gre et al. recently examined various LSTM variants and their relative performances compared to the vanilla model. The vanilla LSTM is commonly understood as the conventional LSTM block, which has been enhanced by including the forget gate and peephole connections.

A total of eight variants were identified for experimentation. In summary, the vanilla architecture demonstrates strong performance across multiple tasks, and none of the eight examined variants exhibit a significant advantage over the others. Many applications described in the literature justify using the standard Long Short-Term Memory (LSTM) model.

In a recent study by Yu et al., a comprehensive examination of the LSTM cell is presented, including its functionalities and various architectural designs. There exists a differentiation between neural networks that are predominantly LSTM-based and those that incorporate LSTM along with other components to exploit its properties, thereby potentially creating hybrid neural networks. Given the inherent uniqueness of each problem, it is frequently advantageous to explore alternative solutions rather than relying solely on the conventional LSTM model (Yu et al., 2022).

The LSTM model, as described in Van Houdt et al. (2020), is a recurrent neural system specifically developed to address the challenges of exploding or vanishing gradients that often occur when attempting to learn long-term dependencies. This is particularly relevant even in cases where the time lags between events are significantly prolonged, as Van Houdt et al. (2020) mentioned.

The composition of a vanilla LSTM unit consists of a cell, an input gate, an output gate, and a forget gate. The forget gate was not included initially in the LSTM network; however, Gers et al. introduced it as a means for the network to reset its state. The cellular system retains information

indefinitely, and the three gates control the flow of information linked to the cell. In the subsequent portion of this section, the term LSTM will denote the conventional version, as it is widely recognised as the predominant LSTM structure (Van Houdt et al., 2020).
Nevertheless, it should be noted that this does not necessarily indicate its superiority in all circumstances.
In brief, the Long Short-Term Memory (LSTM) architecture comprises a collection of interconnected sub-networks called memory blocks. The concept underlying the memory block is to sustain its state over some time and control the flow of information through nonlinear gating units. Figure 1 below illustrates the structural composition of a standard Long Short-Term Memory (LSTM) block. This configuration encompasses the presence of gates, the input signal denoted as x(t), the output signal represented as y(t), and the activation functions. The study by Greg Van Houdt and his colleagues (Van Houdt et al.) explored the concept of peephole connections [32]. The output of the block is connected recurrently to both the block input and all the gates.

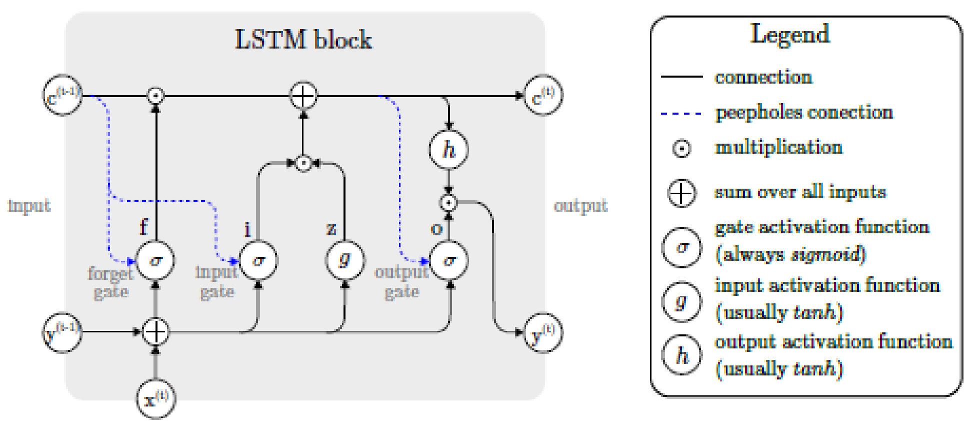


Figure 2.1: LSTM Model

2.5.2. Why LSTM is superior to RNN

Because of the additional complexities introduced by the gates, it may be challenging to understand precisely why the LSTM is superior to the RNN. Unlike RNN, an actual memory is incorporated into the design of an LSTM, which does not have this feature. Because of how we update the cell memory with new information through addition, the LSTM will always have the same error when it needs to be backpropagated at depth. The addition, rather than determining the subsequent cell state by multiplying its existing state with the new input, stops the gradient from bursting or disappearing entirely. The succeeding cell state is determined by multiplying its current state with the new input (Mikami, 2016). (Even if we must multiply the forget gate to access the memory cell.)

2.5.3. Brief Overview of Training, Testing, and Validation

Supervised and unsupervised learning difficulties can be categorised under "learning problems." The three types of supervised learning are classification, prediction, and regression. In each of these types of learning, the input vectors are matched with their corresponding target vectors. The objective here is to predict the output vectors using the input vectors as a basis. Clustering is an example of unsupervised learning. There are no target values in this type of learning; instead, the objective is to explain the correlations and patterns among a set of input vectors (Olah, 2015). A supervised learning issue is solved by the two different implementations of the LSTM algorithm. The challenge states that, given a series of inputs, we wish to predict the probability of the next output.

When performing machine learning, it is best practice to separate a given dataset into three parts: a training set, a validation set, and a test set. These sets are called the training, validation, and test sets. The training set is utilised for learning, the validation set is used to estimate the prediction error for model selection, and the test set is utilised to assess the generalisation error of the ultimately selected model. A good rule of thumb is to allocate fifty per cent of the dataset for training, followed by validation and testing at 25 and 25 per cent, respectively (Mikami, 2016). The validation set is utilised to tweak the network parameters depending on the error rate, whereas the test set is not. This distinction is what differentiates the validation set from the test set. Following the validation, we select the model with the highest success level. Because the purpose of the test set is solely to evaluate the effectiveness of the model chosen, tuning cannot take place while the system is being put through its paces.

This is a generative model, meaning we may generate new text or music by sampling from the output probabilities. In this scenario, the objective is to construct a model capable of predicting the following word or musical note. Because of this, training sets and validation sets are generated to fine-tune the model. A test set is not generated. Instead, a randomly selected batch of data is fed from the training to construct an output. This enables the generation of the output.

There is the possibility that the mistake rate during training will be relatively low; however, the error rate during testing will be significantly greater. The term for this type of event is "overfitting." The test error, also known as the generalisation error, is the mistake anticipated to be produced by the model when applied to records that have not been seen before (Nielson, 2015). To get desirable outcomes during training, there may be a temptation to make the model more complicated. Nevertheless, there are situations where increasing the complexity does not automatically result in a model that can generalise successfully to test examples. A good model should have relatively low levels of error throughout the training phase and the generalisation phase. Underfitting of a model can also happen if the model has not yet learned the proper structure of the data. In this scenario, the training and generalisation errors are typically significant. Underfitting can occur for several reasons, the most common of which are insufficient data or a model that is not sufficiently complicated. Training neural networks is more of an art than a science, even though there exist theories that try to address potential problems with these solutions.

2.6. Decision Tree Classifier

Tijo and Abdulazeez (2021) explained that one of the commonly employed methodologies in data mining involves utilising systems that generate classifiers. Classification algorithms in data mining can process and analyse large quantities of data effectively. This method allows for formulating inferences about categorical class designations, categorising information based on training sets and class labels, and classifying recently acquired data (Tijo and Abdulazeez, 2021). The field of machine learning encompasses a variety of classification algorithms, and this study examines explicitly the decision tree algorithm. Figure 2.2 depicts the structural composition of a decision tree (DT). Decision trees are widely employed in diverse domains, including but not limited to machine learning, image processing, and pattern recognition (Sathiyanarayanan et al.,2019). Decision Trees (DT) are a sequential model that effectively and cohesively combines a series of basic tests, wherein a numeric feature is compared to a threshold value in each test (Sharma and Kumar, 2017). Constructing conceptual rules is relatively more straightforward than determining the numerical weights in the neural network's interconnections between nodes (Carleo et al.,2019). The primary purpose of utilising DT is for categorization.

Furthermore, decision tree (DT) is a commonly employed classification model in Data Mining (Sharma and Kumar, 2017). The components of each tree consist of nodes and branches. In the context of classification, it can be observed that each node represents distinct features within a given category. Furthermore, it is noteworthy that each subset within this framework delineates the potential values that can be assumed by the respective node (Dey, 2016). Decision trees have been widely implemented in various fields due to their straightforward analysis and accurate handling of multiple data formats (Mrva et al., 2019). Figure 2.2. presents an illustrative example of Decision Trees (DT).

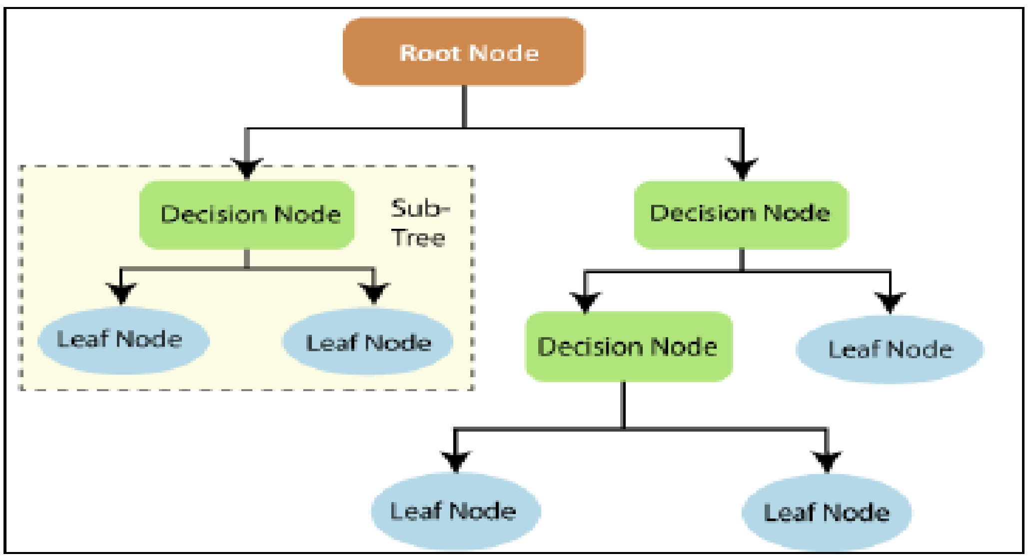


Figure 2.2: Decision Tree

Decision tree learning is a widely employed technique in data mining. The objective is to develop a predictive model that estimates the value of a target parameter by considering multiple input parameters. Training a tree involves partitioning the original dataset into subsets using attribute value tests (Kumar, 2013). As mentioned, the procedure is iteratively applied to every derived subset, following a recursive approach called recursive partitioning.

The recursion process is considered complete when the subset of data at a particular node possesses uniform values for the target variable or when further data splitting no longer contributes any additional value to the predictions. The abovementioned approach of top-down induction of decision trees exemplifies a greedy algorithm, which is widely employed as the predominant strategy for acquiring decision trees from data mining data. In TDIDT systems, machine learning can be categorised based on the following factors (Kumar, 2013). These include the learning strategy employed, the representation of the knowledge acquired by the system, and the system's application domain.

Decision trees are a fusion of mathematical and computational methodologies utilised to describe, categorise, and generalise a specific dataset, thereby facilitating machine learning. In the context of decision trees, the dependent variable is predicted based on the independent variable. Decision tree types commonly employed in data mining are typically categorised differently. Classification tree analysis is a widely used method for predicting data within a specific class. Regression tree analysis is a necessary method for predicting the value of an independent variable, such as the price of a house or the length of stay for a patient in a hospital.

Xiang et al. (2020) explored various research on predicting the weather using the Decision Tree Model of Machine Learning. He explained that the decision tree model is an outstanding example of an algorithm for machine learning. It uses recursive segmentation technology to continuously divide the data space into multiple subsets to identify the possible structure, significant patterns, and relationships within the data. Compared to conventional parametric statistical methods, the decision tree model does not require any prior assumptions to be made on the relationship between independent factors and dependent variables, and it is also able to address the problem of multi-collinearity among independent variables. On the other hand, the outputs of a single decision tree are prone to overfit and are highly unstable. Using Bootstrap sampling technology, the random forest model constructs decision trees by randomly taking some samples from the original. It then merges numerous decision trees to avoid overfitting successfully (Zhao, 2017). At the moment, decision trees and random forest algorithms are being utilised in meteorology increasingly widely. In another work reviewed, the decision tree method established a more accurate categorization and prediction model for the road icing disaster. The classification and prediction model of regional summer precipitation days was examined by (Shi Yimin et al.2018) based on the data mining Classification and Regression Tree (CART) algorithm. Qin Pengcheng et al. (2016) demonstrated that an analysis of Hubei rapeseed yield-limiting factors based on a decision tree and a random forest model was successful in its application.

3. METHODOLOGY

3.1 Research Methodology

This section outlines the sequential measures to address research techniques and achieve study objectives. The Knowledge Discovery in Databases (KDD) approach was utilised for this study analysis, aiming to extract crucial knowledge, patterns, and high-level information from relevant datasets stored in databases. Databases serve as valuable tools for generating and validating information from various perspectives, such as information management, query processing, decision-making, and process control, among other applications that can benefit from the insights gained through this method.

Data mining techniques are becoming increasingly essential with the growth of information services applications like the World Wide Web and online services since it enables researchers to gain deeper insights into user behaviour while improving service quality and seizing potential commercial opportunities. The KDD approach involves data integration and selection as its initial two phases, followed by data cleaning and preprocessing before transitioning into data transformation, mining, evaluation, and interpretation processes.

This systematic approach allows researchers to extract valuable information from databases, maximising its potential use across various fields and applications.

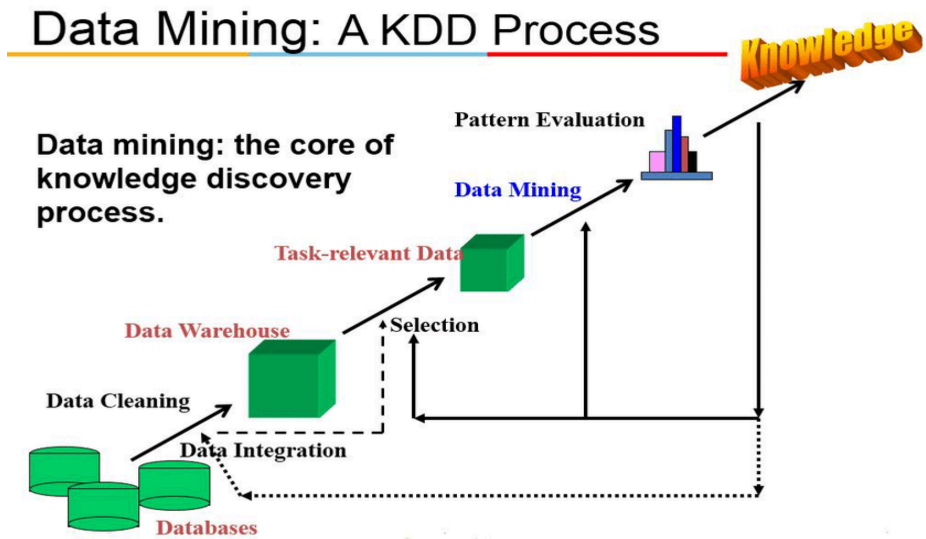


Figure 3.1: The KDD Lifecycle

This section will thoroughly explain the approach, reviewing each step in depth. It will cover the steps involved in gathering and choosing data, cleaning it, and preparing it for analysis. Additionally, the data will be transformed to enable efficient analysis. We will use exploratory data analysis (EDA), a crucial step in the data mining process, to draw patterns and conclusions from the processed data.

Some of the analyses to be done include;

a) Feature engineering of the datasets
b) Training a deep learning model for the temperature data prediction using LSTM and Decision Tree Classifier for the UV Radiation Prediction
c) Analyse the model's performance and, where applicable, finetune the model to perform better.
d) Wrap the different solutions into two functions: function one will predict the next temperature values given the current or previous values, and function two will predict the UV Radiation, given the temperature.

a) The datasets underwent an extensive feature engineering process, consisting of variables being modified, merged, and extracted to produce improved input features for the subsequent modelling stages. During this stage, the objective was to transform the raw data into a format allowing a more accurate representation of the underlying relationships and patterns, enhancing the models' capacity to make accurate predictions.

b) The development of predictive models that could accurately forecast temperature and ultraviolet radiation was the primary focus of this project. A deep learning model was utilised to analyse the temperature data, which was chosen because of its capacity to recognise complex patterns within the dataset. A robust method known for partitioning data into various classes was responsible for simultaneously handling the UV radiation prediction. This algorithm was a Decision Tree Classifier. These two unique methods were used because each prediction job benefited more from one than the other.

c) A thorough investigation into the performance of the models was carried out, which involved utilising a variety of measures such as Mean Squared Error, Accuracy, and Precision. This examination made gaining a well-informed view of the models' prediction capabilities possible by drawing attention to their strengths and potential weaknesses. A process of fine-tuning was carried out whenever there was space for improvement in the models that were being used. During this phase, adjustments were made to the hyperparameters, model topologies, and feature selections. These changes improved the model's predicted accuracy and overall performance.

d) To encapsulate the produced solutions, a strategic approach was chosen to encapsulate the models into functional units. This was done to accomplish the goal of encapsulating the solutions. The first function's objective was to forecast future temperature readings by extrapolating from the most recent or most recent measurements. The function created accurate temperature forecasts by utilising the trained deep learning model and consuming these inputs, which aided in future planning and decision-making. The second function, on the other hand, focused explicitly on forecasting UV radiation. This function used a trained Decision Tree Classifier to forecast UV radiation levels. It did so by receiving temperature data as an input and using it to predict UV radiation levels. This made it easier to assess potential exposure hazards and take safety precautions.

The project was developed via separate phases, including data transformation, model training, performance analysis, and solution encapsulation in its most basic version. This systematic process ensured a solid approach for predictive analysis, which offered helpful insights into variations in temperature and UV radiation and made it possible for informed measures to be taken in response to changing environmental conditions.

3.2. Data Selection and Integration

The data for the study were taken from the Kaggle website, an online storage of many data kinds applicable to various domains. The data source is linked at:
https://www.kaggle.com/datasets/deepcontractor/mars-rover-environmental-monitoring-station.

For better results, this project will use one of two datasets. This project would be using one of two datasets for better results. Details about each dataset's summary are provided below.

1st Dataset

The dataset tagged mars-weather-1 is saved in CSV format, which has data on weather conditions on Mars starting from August 7, 2012 (represented as Sol 1) to February 27, 2018 (represented as Sol 1895). The Rover Environmental Monitoring Station (REMS), which is mounted on the Curiosity Rover, measured and transmitted the data (more details about REMS can be found at https://mars.nasa.gov/msl/spacecraft/instruments/rems/)

[Rover Environmental Monitoring Station (REMS) is a weather station on Mars for the Curiosity rover, contributed by Spain and Finland. REMS measures humidity, pressure, temperature, wind speeds, and ultraviolet radiation on Mars. The Spanish Astrobiology Center leads this Spanish project and includes the Finnish Meteorological Institute as a partner, contributing pressure and humidity sensors. source: !https://www.kaggle.com/datasets/deepcontractor/mars-rover-environmental-monitoring-station]

Attributes Description

1. id - The identification number of a single transmission
2. terrestrial_date - The date on Earth (formatted as month/day/year or m/dd/yy).
3. ls - The solar longitude or the Mars-Sun angle, measured from the Northern Hemisphere. In the Northern Hemisphere, the spring equinox is when ls = 0. Since Curiosity is in the Southern Hemisphere, the following ls values are of importance: • ls = 0: autumnal equinox • ls = 90: winter solstice • ls = 180: spring equinox • ls = 270: summer solstice
4. month - The Martian Month. Similarly, to Earth, Martian time can be divided into 12 months.
5. min_temp - The minimum temperature (in °C) observed during a single Martian sol.
6. max_temp - The maximum temperature (in °C) observed during a single Martian sol.
7. pressure - The atmospheric pressure (Pa) in curiosity's location on Mars.
8. wind_speed - The average wind speed (m/s) measured in a single sol. Note: Wind Speed data has not been transmitted to Earth since Sol 1485. Missing values are coded as NaN.
9. atmo_opacity - Description of the overall weather conditions on Mars for a given sol based on atmospheric opacity (e.g., Sunny)

2nd Dataset

The weather data in this second collection, mars-weather-2, which is kept in comma-separated values (CSV), offers details about the weather on Mars from 2012 to 2022. The dataset will be utilised instead of the earlier one because it has more recent data. The details of Mars's time and weather are described at: https://www.planetary.org/articles/mars-calendar

The following procedures would be used to prepare the data as well:

1. Installing all required libraries.
2. Loading the dataset into the pandas environment.
3. Spotting missing values and dealing with them appropriately.

The preliminary phase of the research project encompassed establishing the requisite computational infrastructure, compilation and organisation of the data, and implementation of thorough data preprocessing, adhering to established scholarly protocols.

Installing all Required Libraries*:* The initial phase of the research methodology involved the installation of necessary software libraries and adhering to established principles of sound research. The libraries mentioned are designed to cater to the unique requirements of the investigation, offering a structured approach for the manipulation, analysis, and visualisation of data. The installation of these libraries by researchers ensured the availability of a comprehensive range of tools that facilitated the efficient management of subsequent data handling procedures.

Loading the dataset into the pandas environment: The subsequent phase of the study focused on loading the research dataset into the Python programming environment, utilising the versatile "pandas" library on the jupyter notebook IDE. The execution of this crucial stage was executed meticulously, as the precise representation of data forms the foundation for any empirical analysis. Researchers successfully utilised the functionalities of the "pandas" library to convert the unprocessed data into an organised format, facilitating subsequent phases of investigation and examination.

Spotting missing values and dealing with them appropriately: The identification and management of missing values is a crucial component of thorough data analysis. The diligent research methodology of methodically examining the dataset uncovered instances in which values were missing. It was imperative to address these gaps diligently, as incomplete data can distort analyses and undermine the credibility of research findings. The researchers demonstrated methodological rigour in their approach to handling missing values, employing techniques such as imputation, which involves estimating missing values based on available data, or deletion, which consists of removing incomplete records. The importance of this step was emphasised through a careful evaluation of its potential effects on the integrity of the dataset and the validity of subsequent analyses.

In summary, the initial phases of the research process highlighted the diligent and meticulous approach to scholarly investigation. The commitment to methodological rigour was demonstrated through the installation of libraries, loading of data, and careful handling of missing values. These actions established a strong foundation for subsequent analyses and findings. The strict adherence to established best practices in data preparation and preprocessing played a significant role in upholding the scholarly integrity and credibility of the research endeavour.

These procedures guarantee that the data is in an appropriate format for additional analysis and modelling. Dealing with missing information can prevent problems resulting from incomplete or erroneous data.

The python environment, which was utilised for development, had important libraries installed for data wrangling and visualisation, preprocessing, and modelling. Libraries like Pandas, Numpy, and Matplotlib were imported for data manipulation and visualisation. Additionally, deep learning frameworks such as TensorFlow were imported into the environment. For consistency of results, tf.random.set_seed was instantiated. A Windows….The two datasets described above were loaded into the environment, and the data were previewed.

3.3. Data Cleaning and Preprocessing

Data cleaning and preprocessing is the first or primary step of data analytics, used to interpret, examine, and understand the data to quickly gain crucial insights or identify trends or important areas to investigate further (Joshi and Patel, 2022). Charts, reports, and other visuals, along with human and automatic tools, are used. This enables us to gain the most knowledge from the possible data, identify any outliers or erroneous data and abnormalities, test underlying theories, and select the most appropriate factor settings.

After generating the intended data, data preprocessing is necessary to make it usable. The ready-to-use data is used in the following stage to analyse the data and produce information or results using various mining techniques. Data cleaning, data optimisation, data transformation, data integration, and data conversion are five of the data preparation approaches.

Data visualisation and numerical summary are powerful tools for exploring data and potent methods of communicating conclusions. In the preprocessing stage of the data mining process, visualisation techniques play a vital role in identifying inaccurate values, missing values, duplicate rows, columns with the same value throughout, and other issues that aid in data cleansing.

For this study, the first data was previewed using data1.head() method and the same was done for the second data. The data tail was also viewed.

3.4.1. Data Wrangling

Teng and Rong (2022) explained that data wrangling, also known as data munging, encompasses the activities involved in converting and aligning data from its original, unprocessed state to a different format, aiming to enhance its suitability and usefulness for various subsequent applications, such as analytics.

It was observed that the second dataset differs from the first in how missing values were represented, suggesting that the first dataset may have undergone some preprocessing. To remedy the missing data representation, string replacement was used. From the data description, wind speed does not have data since sol 1485 because it was not transmitted. The replace function was applied, and the result was previewed.

	earth_date_time	mars_date_time	sol_number	max_ground_temp(°C)	min_ground_temp(°C)	max_air_temp(°C)	min_air_temp(°C)	mean_pressure(Pa)
3187	Earth, 2012-08-23 UTC	Mars, Month 6 - LS 159°	Sol 17	-4	-76	7	-81	742
3188	Earth, 2012-08-22 UTC	Mars, Month 6 - LS 158°	Sol 16	0	-77	9	-81	740
3189	Earth, 2012-08-21 UTC	Mars, Month 6 - LS 158°	Sol 15	-15	-78	8	-82	740
3190	Earth, 2012-08-20 UTC	Mars, Month 6 - LS 157°	Sol 14	-16	-74	9	-82	740
3191	Earth, 2012-08-19 UTC	Mars, Month 6 - LS 157°	Sol 13	-15	-74	8	-80	732
3192	Earth, 2012-08-18 UTC	Mars, Month 6 - LS 156°	Sol 12	-18	-76	8	-82	741
3193	Earth, 2012-08-17 UTC	Mars, Month 6 - LS 156°	Sol 11	-11	-76	9	-83	740
3194	Earth, 2012-08-16 UTC	Mars, Month 6 - LS 155°	Sol 10	-16	-75	8	-83	739
3195	Earth, 2012-08-15 UTC	Mars, Month 6 - LS 155°	Sol 9	NaN	NaN	NaN	NaN	NaN
3196	Earth, 2012-08-07 UTC	Mars, Month 6 - LS 150°	Sol 1	NaN	NaN	NaN	NaN	NaN

Figure 3. 2: Second dataset Tail

Unique values in the identified column were previewed. It was observed that the wind speed and humidity columns would be removed from the tables due to preliminary observations that they do not contain any information that will benefit this analysis. Secondly, the second data will be used for this research because the second dataset provides considerably more recent information about Mars. Going with the new dataset, the columns that were not needed were dropped.

	earth_date_time	mars_date_time	sol_number	max_ground_temp(°C)	min_ground_temp(°C)	max_air_temp(°C)	min_air_temp(°C)	mean_pressure(Pa)	sun
0	Earth, 2022-01-26 UTC	Mars, Month 6 - LS 163°	Sol 3368	-3	-71	10	-84	707	0
1	Earth, 2022-01-25 UTC	Mars, Month 6 - LS 163°	Sol 3367	-3	-72	10	-87	707	0
2	Earth, 2022-01-24 UTC	Mars, Month 6 - LS 162°	Sol 3366	-4	-70	8	-81	708	0
3	Earth, 2022-01-23 UTC	Mars, Month 6 - LS 162°	Sol 3365	-6	-70	9	-91	707	0
4	Earth, 2022-01-22 UTC	Mars, Month 6 - LS 161°	Sol 3364	-7	-71	8	-92	708	0

Figure 3.3: The data after columns were dropped

3.4.2. Preprocessing the new Data

For better isualizing and modelling, the new data will be processed to reflect a structure resembling the old one. The Earth's date time was used to extract the datetime from strings while solar longitude, which is a measurement of where the planet is in relationship to the sun, was extracted from the Mars date time and also to create a new column called "Mars month". Then, the solar number was used to extract the integer representation of the solar expedition, which, starting with the mission's first day, represents each Martian day. All of these would be achieved using the regular expression library in Python, like using pandas to convert time.

	earth_date_time	mars_date_time	sol_number	max_ground_temp(°C)	min_ground_temp(°C)	max_air_temp(°C)	min_air_temp(°C)	mean_pressure(Pa)	sunrise	sunset	UV_Radiation	weather	month
0	2022-01-26	163	3368	-3	-71	10	-84	707	05:25	17:20	moderate	Sunny	6
1	2022-01-25	163	3367	-3	-72	10	-87	707	05:25	17:20	moderate	Sunny	6
2	2022-01-24	162	3366	-4	-70	8	-81	708	05:25	17:21	moderate	Sunny	6
3	2022-01-23	162	3365	-6	-70	9	-91	707	05:26	17:21	moderate	Sunny	6
4	2022-01-22	161	3364	-7	-71	8	-92	708	05:26	17:21	moderate	Sunny	6

Figure 3.4: Renamed Data

As seen above, some columns were renamed or rearranged to ease understanding.

3.4.3. Handling Missing Data

The missing data were first viewed across the whole dataset. The total length of the whole dataset was seen as it was observed that earth date, Ls , sol, month, sunrise and sunset were the only columns with no missing data and max_ground temp, min_ground _temp, max_air_temp, min_air_temp, mean_pressure, UV_radiation and weather has 28,28, 29,29, 27, 27 and 3 missing data respectively from a total data frame length of 3,197. The missing data is 0.9%, which is not significant; hence, the missing data will be dropped.

3.4.4. Further Data Wrangling

- The following operations would be done using the data:

    1. Calculate the duration of sunlight time across the data.
    2. Calculate the seasons given the guide from the image attached.
    3. Calculate the number of years covered in the dataset using the provided source.

- Make MARS season calculations utilising the longitudinal solar data.

    - According to the information below, the Spring falls between 0º and 89º ls, and so on.

Areocentric longitude (L_s)	Season	
	Northern Hemisphere	Southern Hemisphere
0° to 89°	Spring	Autumn (Fall)
90° to 179°	Summer	Winter
180° to 269°	Autumn (Fall)	Spring
270° to 359°	Winter	Summer

Figure 3.5: Data Image

    - Add a new season-related column.

The unique values in the season column were viewed, and it was observed that, given that the planet's yearly calendar is nearly two times as long as the calendar of planet Earth and that the sol column begins at 10 in the Earth year of 2012, these facts are also relevant. The study then used the suggestion from (https://www.planetary.org/articles/mars-calendar) to create a Mars year column to capture the years in the data; here is how the years on Mars are distinguished.

1. Year 31 is denoted by the numbers 0 to 350,
2. Year 32 by the numbers 351 to 1018,
3. Year 33 by the numbers 1019 to 1687,
4. Year 34 by the numbers 1688 to 2356,
5. Year 35 by the numbers 2357 to 3023
6. Year 36 by the numbers 3024 to 3368.

- Make a column for mars_year.

The mars_year filter was applied, and the data was visualised more for insight.

3.5 Visualising the Data for More Insights

It is essential to illustrate the weather patterns for various temperatures (air and ground temperature) and the UV radiation intensity that can be experienced at any given moment, given that this project aims to determine the likelihood of man existing on Mars. As a result, it is imperative to visualise the different weather patterns for different temperatures (air and ground temperature), and the UV radiation level obtainable per time. So therefore, the visuals would include the temperature and UV radiation columns.

Boxplots will be used as they help identify the presence of outliers in a dataset, which can then be used to decide how to treat them earlier in the research process. Outliers can serve to highlight anomalous occurrences or errors in the data.

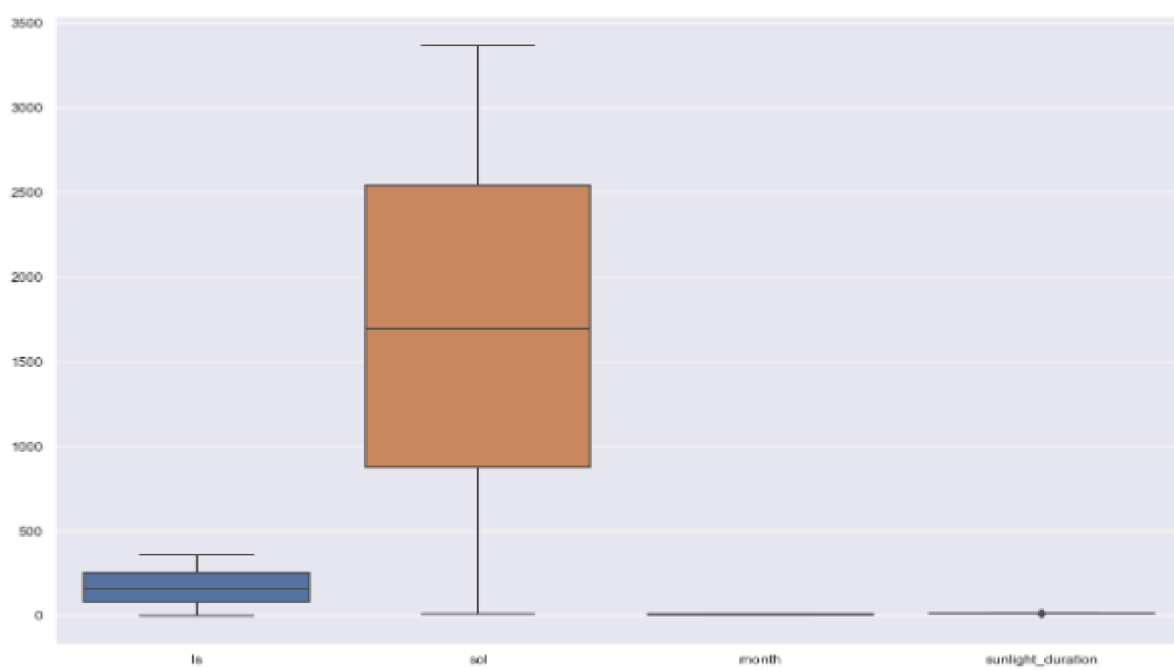


Figure 3. 6: Boxplot for Outliers

The data has no apparent outliers, according to the plot above. Since there are no outliers, the data are appropriate for the analysis that will be done.

3.5.1. Temperature

The plot data was selected on the x-axis, while the y-axis data was selected. A function to plot the data was selected as the density was presented in the data set, as seen in Figures 3.6 A and B

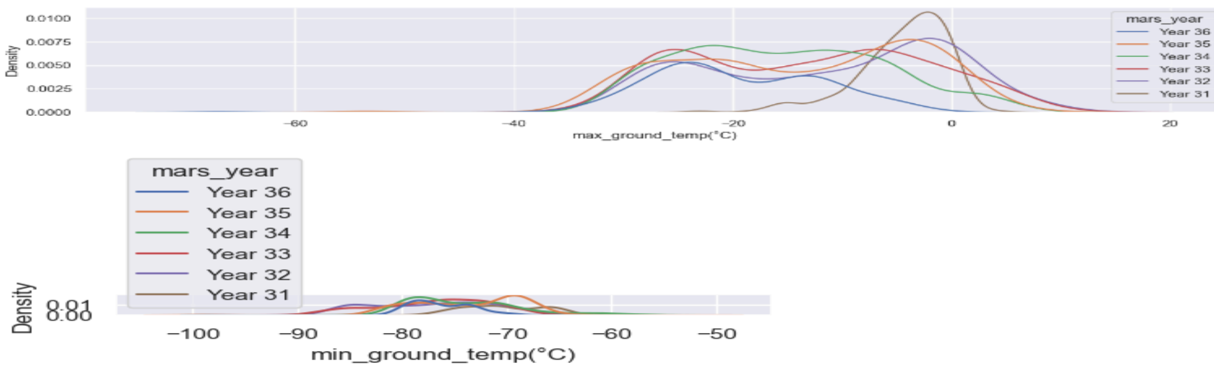


Figure 3.7 A: A Density plot showing the data spread of max_ground_temp(C) across the mars year column.

Figure 3.7 B: A Density plot showing the data spread of min_ground_temp(C) across the mars year column.

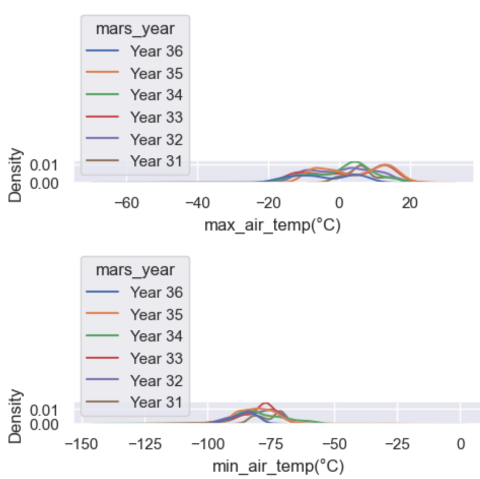


Figure 3.8 A: A Density plot showing the data spread of max_air_temp(C) across the mars year column.

Figure 3.8 B: A Density plot showing the data spread of min_air_temp(C) across the mars year column.

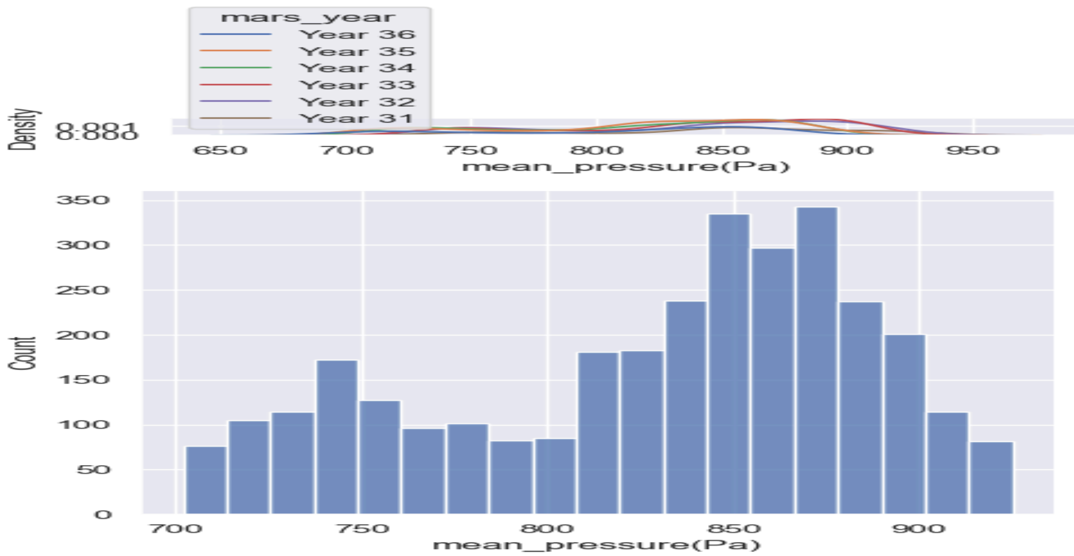


Figure 3.9 A: A Density plot showing data spread of the mean_pressure(C) across different mars years.

Figure 3.9 B: A bar chart showing more detailed data distribution across different variables.

The first plot (Fig 3.7 & 3.8) shows the intersection of the dataset's highest ground temperatures measured throughout various Mars years. It demonstrates that the 31st Mars year recorded the greatest maximum ground temperature. However, the maximum ground temperatures in other years showed a more consistent pattern without notable peaks or changes.

The mean pressure readings that were recorded on Mars are shown in the second plot (Fig 3.9). It demonstrates that the pressure on Mars varies from about 830 to about 900 on average. This pressure range seems to be the most common one found on Earth.

The figure elucidates the usual atmospheric pressure conditions on Mars and aids in our comprehension of the dataset's mean pressure readings. The full plot is presented in Figure 3.10 below.

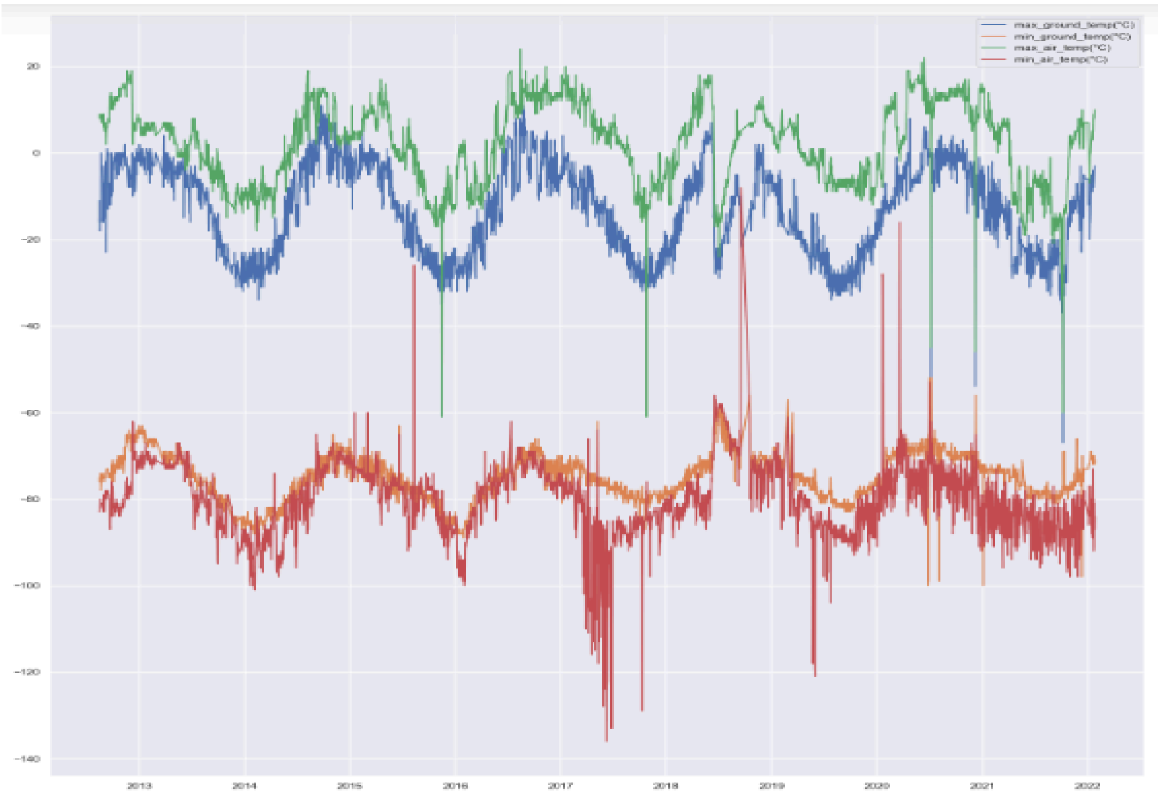


Figure 3.10: A plot showing the relationship between dates and different temperatures.

The graphic shows a wave-like pattern representing the correlation between dates and the chosen ground and air temperatures on Mars. This trend implies that the changes in global temperature are subject to regular seasonality. The increase and decrease in temperatures over time suggest that Mars' climate follows cyclical patterns, much like the seasons on Earth.

It is now conceivable to forecast the anticipated temperature on Mars based on the date or season, thanks to this established pattern of seasonality. We may predict temperature trends and make accurate assumptions about the planet's climate by comprehending the cyclical pattern of temperature variations. This knowledge may be helpful when organising and carrying out missions or researching the Martian environment.

3.5.2. UV Radiation

UV Radiation is emitted by the sun, which is a higher frequency and lower wavelength than visible light, so it can't be seen with the human eyes, and overexposure could have serious health implications. UV Radiation has a higher frequency and shorter wavelength than visible light; ultraviolet (UV) radiation has a wavelength range of 100–400 nm. The sun is a natural source of UV radiation, but it can also be produced artificially by sources utilised in business, industry, and leisure.

Three bands make up the UV area, which has wavelengths between 100 and 400 nm: UVA (315-400 nm), UVB (280-315 nm), and UVC (100-280 nm). Ozone, water vapour, oxygen, and carbon dioxide absorb all UVC and around 90% of UVB radiation as it travels through the atmosphere. The atmosphere has little impact on UVA radiation. As a result, UVA dominates the UV radiation that reaches the Earth's surface, with a modest amount of UVB radiation.

Better summarised in the format below.

- UVA (315-400 nm)
- UVB (280-315 nm)
- UVC (100-280 nm).

The study proceeded to investigate UV Radiation

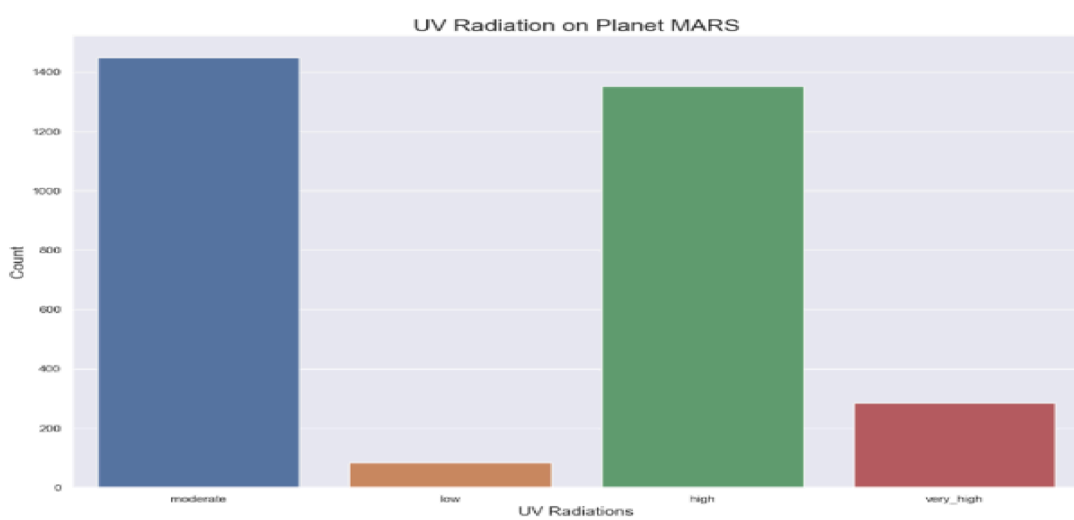


Figure 3.11: UV radiation Barchart

According to the analysis, Martian UV radiation levels are typically moderate. Many of the data points show a more than 40% chance of encountering moderate UV radiation at any particular time in the future. The level of UV radiation can be high at times; therefore, it's vital to remember that this may not always be the case.

High UV radiation levels suggest Mars occasionally encounters elevated amounts of UV radiation, even if moderate UV radiation levels seem more prevalent. Environmental elements, including solar activity, air conditions, and other ecological variables, may impact these high levels.

Therefore, the likelihood of experiencing high UV radiation levels on Mars must be considered, even though moderate UV radiation levels predominate the dataset. To ensure the safety and well-being of humans or equipment during trips to Mars, it is essential to comprehend the fluctuations in UV radiation and its potential effects.

3.5.3. Solar Longitude, Seasons and Sunlight

Plotting the relationship between solar longitude and sunlight duration as the seasons

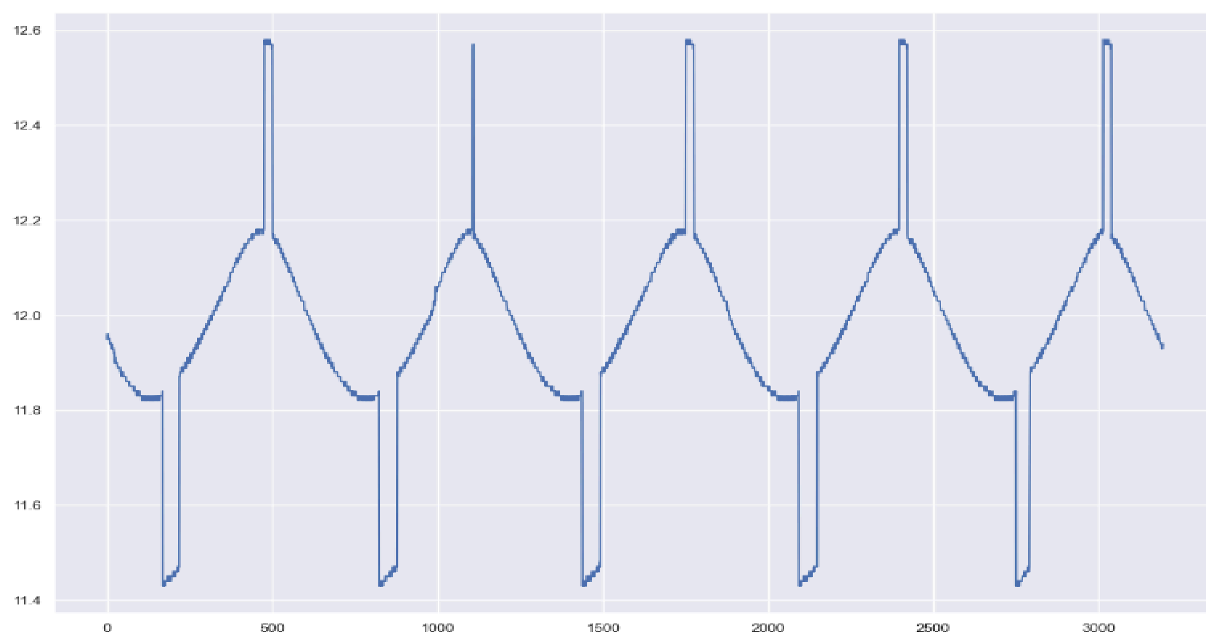


Figure 3.12: A plot showing the relationship between solar longitude, sunlight duration

The upside-down funnel-shaped graphic shows how the length of sunshine on Mars remains constant over the years. This pattern suggests that the amount of sunlight accessible on the planet is steady and predictable. Given the narrow funnel form, it is likely that Mars's sunshine duration is relatively stable and does not undergo significant variations or dramatic fluctuations.

The predictability of sunlight duration on Mars is noteworthy compared to Earth, where it can vary considerably due to elements including latitude, season, and weather patterns. The plot's recurring pattern shows that Mars experiences a roughly constant amount of sunshine across time, making its availability more predictable than on Earth.

Planning missions, creating solar-powered equipment, and researching Mars' surface all benefit from understanding the predicted length of the planet's sunlight. Better resource management is made possible by it, and offers an understanding of the daily and seasonal fluctuations in solar energy on Mars.

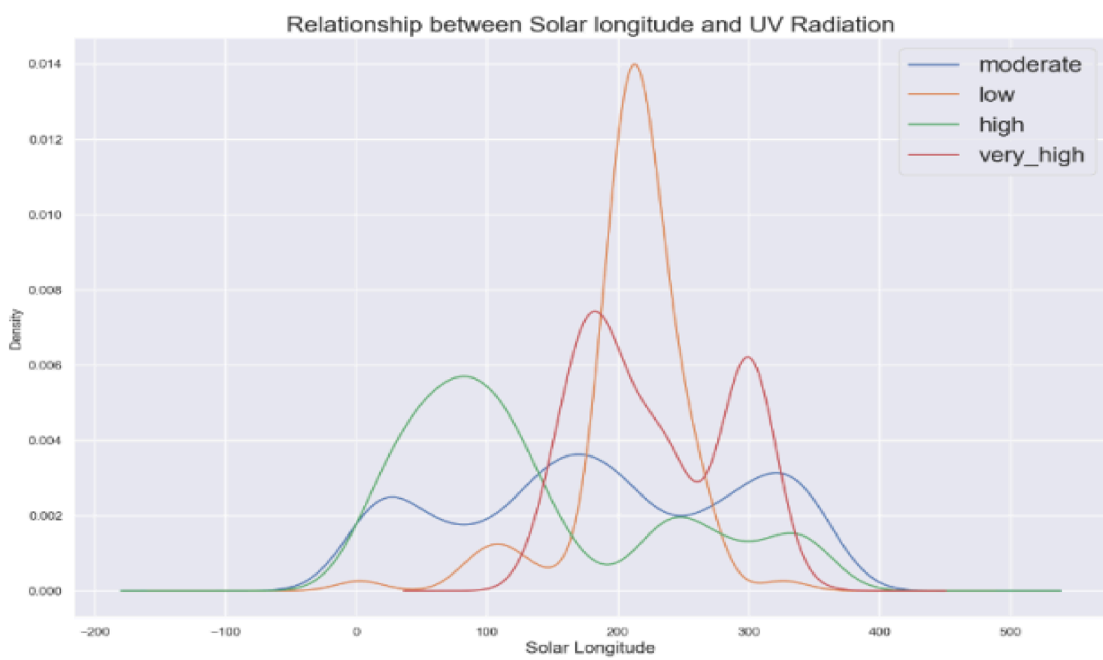


Figure 3.13: Relationship between solar Longitude and UV Radiation

The graphic demonstrates abrupt increases in very high and low UV radiation levels on Mars during sols 100 and 300 before each level, depicting a steady pattern towards the end of the plot. The UV radiation levels had a more regular pattern before these surges. These unexpected surges point to unusual occurrences or events that might have happened during those longitudinal solar cycles.

There may be a few causes for these unusual increases in UV radiation levels. It might be connected to the regional atmosphere, specifically dust storms or cloud cover, which can affect how much UV light reaches the surface. These surges might potentially be attributed to solar flares or other solar events.

To better comprehend the reasons behind these spikes and any potential ramifications, it is crucial to investigate the incidents or conditions that gave rise to them. The dynamic nature of Mars' atmosphere and the elements affecting its radiation environment can be better understood by tracking and evaluating UV radiation levels for various sol periods.

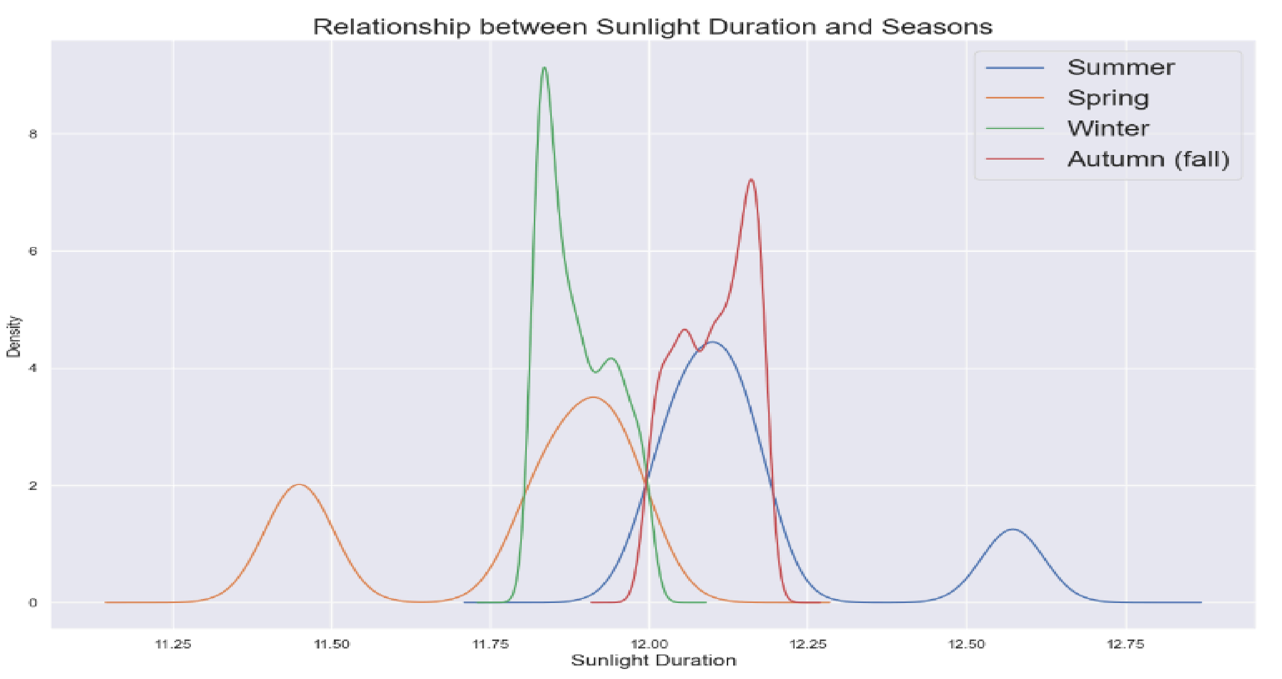


Figure 3.14: Relationship between Sunlight Duration and Seasons.

According to the figure, Mars' winter and fall sun exposure times fall within a predicted range of 11.75 to 12.25 hours. This continuous range implies that Mars's sunshine duration is mainly predictable and steady during these seasons. In terms of the amount of sunlight accessible in winter and fall, the small range denotes a high degree of predictability.

Additionally, it's significant that all four seasons coincide with 12 hours of sunlight. This intersection suggests that it is possible to have 12 hours of sunlight on Mars throughout the year. This shows a constant baseline of 12 hours of sunlight, regardless of the season.

Planning and carrying out expeditions, conducting agricultural operations, and being aware of the Martian climate depend on knowing the predicted sunshine duration throughout the various seasons on Mars. It enables better resource management and aids explorers and researchers in predicting when sunlight will be available for specific tasks on the planet.

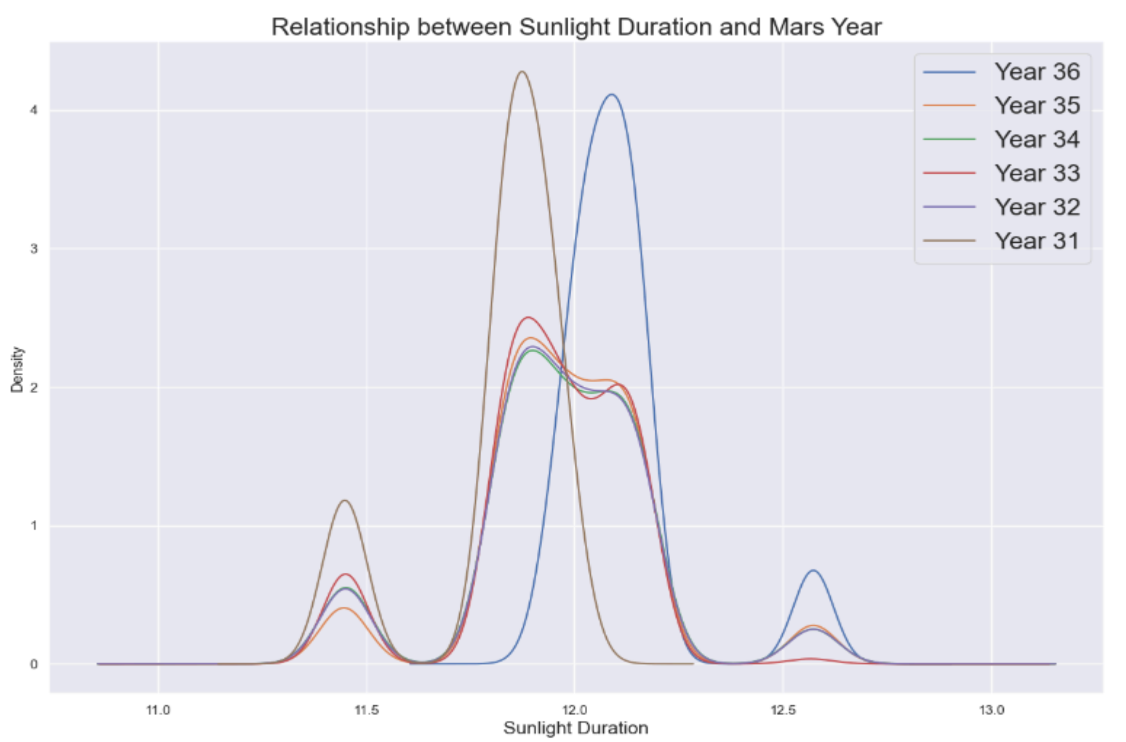


Figure 3.15: Relationship between Sunlight Duration and Mars Year

The graphic shows that most of the dataset's covered years had sunlight durations near 12 hours at ideal levels. This indicates that the amount of sunshine accessible on Mars over these years followed a regular and predictable pattern. Particularly in one year, the length of the day was slightly longer than in the other years, showing a minor fluctuation in the dataset.

The Mars years 31 and 36 stand out as having the longest daylight, though. Compared to other years in the dataset, these years had noticeably longer daylight stretches. The axial tilt of Mars, the planet's elliptical orbit around the sun, and other factors may all be to blame for this variation.

The finding of increased sunlight times in Mars years 31 and 36 emphasises the Martian climate's inherent unpredictability and how it affects the amount of sunlight available. For mission planning, agricultural endeavours, and any other operations that depend on the availability of the sun on Mars, comprehension of these fluctuations is essential.

4. RESULTS

4.1. Prediction Modelling

Mars's ground and air temperatures are crucial for determining if the environment is suitable and comprehending the climate. Given that the ground temperature is a good way to measure the proper temperature that will support life, while the air temperature helps to determine the likelihood of a change in weather conditions, the two will be used to ascertain and predict the temperature on planet Mars.

Data preparation techniques like scaling and preprocessing are required to forecast Mars's temperature. Scaling ensures that the data is uniform and each characteristic has a comparable scale. The data needs to be first scaled, which reduces data weight that has no significant impact on the data and vice versa.

A predictive model based on the correlation between ground and air temperatures can be created after data preparation using modelling techniques like regression or time series analysis. Then, Mars's temperature may be predicted using this model. We can learn more about the Martian environment and its implications for upcoming missions and scientific research by examining temperature patterns.

A new datafarme object is first created containing earth_date, min_ground_temp, max_ground_temp(°C), min_air_temp(°C), max_air_temp(°C). The preprocessing and scaling were defined.

Time series analysis often involves daily resampling, frequently resulting in missing data points (NaN values). The 'df.isna().sum()' method computes the sum of NaN values in each column to determine whether any values are missing. One typical method for handling missing values is filling them with suitable ones. Using the fillna() method with the mean of each column as a replacement for missing values is helpful; however, multiple model finetuning showed that filling with 0s enabled the model to learn better from the dataset. The distribution and properties of the data are preserved by this method. The integrity of the time series is maintained by imputing missing values of this nature with zeros, enabling a more precise analysis. As presented below, the preprocessing method was checked to see if it worked.

```
min_ground_temp(°C)    283
,max_ground_temp(°C)    283
,min_air_temp(°C)       283
,max_air_temp(°C)       283
,dtype: int64
```

Before filling in missing data

```
min_ground_temp(°C)    0
,max_ground_temp(°C)    0
,min_air_temp(°C)       0
,max_air_temp(°C)       0
,dtype: int64
```

After replacing missing values

- The scaler functions to scale the selected data.
- On the other hand, the preprocessor resamples the data and creates the new scaled data by using the scaling function.

This was done using the min-max scalar due to the large difference in the temperature results. The preprocessed daily data is presented below.

```
array([[ 4.1666985e-02,  3.0769238e-01, -1.7187500e-01,  6.2352943e-01],
       [ 2.3841858e-07,  4.3589750e-01, -1.7187500e-01,  6.4705884e-01],
       [ 2.3841858e-07,  2.5641033e-01, -1.5625000e-01,  6.2352943e-01],
       ...,
       [ 2.5000024e-01,  6.1538470e-01, -1.4062500e-01,  6.2352943e-01],
       [ 1.6666698e-01,  6.4102572e-01, -2.3437500e-01,  6.7058825e-01],
       [ 2.0833349e-01,  6.4102572e-01, -1.8750000e-01,  6.7058825e-01]],
      dtype=float32)
```

Figure 4.1: Preprocessed data

The next step was to set the data generator up with a default time step of 1 before applying the scaling. The data was split into train and test data, and a higher look back of about ten was used.

```
array([[ 4.1666985e-02,  3.0769238e-01, -1.7187500e-01,  6.2352943e-01],
       [ 2.3841858e-07,  4.3589750e-01, -1.7187500e-01,  6.4705884e-01],
       [ 2.3841858e-07,  2.5641033e-01, -1.5625000e-01,  6.2352943e-01],
       [ 8.3333492e-02,  3.3333340e-01, -1.2500000e-01,  6.2352943e-01],
       [ 8.3333492e-02,  3.0769238e-01, -1.5625000e-01,  6.4705884e-01],
       [-8.3333015e-02,  3.3333340e-01, -1.5625000e-01,  6.2352943e-01],
       [-4.1666508e-02,  7.1794879e-01, -1.4062500e-01,  6.4705884e-01],
       [ 2.3841858e-07,  6.1538470e-01, -1.4062500e-01,  6.0000002e-01],
       [ 4.1035652e-02,  3.7937072e-01, -1.2996173e-01,  4.8260552e-01],
       [ 4.1035652e-02,  3.7937072e-01, -1.2996173e-01,  4.8260552e-01]],
      dtype=float32)
```

Figure 4.2: Data after Look back

The shape of the train data was seen to be (2749, 10, 4), indicating that the train dataset has 2749 samples, each having ten attributes and each feature having four sub-attributes. This kind of structure is common in multidimensional data.

4.2. Model Architecture Structure

4.2.1. Overview

The neural network architecture consists of visible LSTM hidden layers with 256 blocks (neurons) each, a Repeat Vector that will take the first shape of the data input shape, a

Bidirectional LSTM layer, a DENSE layer with 64 blocks and an output layer that predicts a four-temperature value (minimum and maximum ground-level temperature, minimum and maximum air temperature).

- **Dense -** The regular densely connected NN layer.
  Dense implements the operation: output = activation (dot(input, kernel) + bias) where activation is the element-wise activation function passed as the activation argument, the kernel is a weights matrix created by the layer, and bias is a bias vector created by the layer (only applicable if use_bias is True). (source -> Docstring)
- **RepeatVector - Repeats the input n times.**
- **LSTM -** Implement some hidden layers here, and, based on available runtime hardware and constraints, this layer will choose different implementations (cuDNN-based or pure-TensorFlow) to maximise the performance.
- **Bidirectional -** Bidirectional wrapper for RNNs.
- ***Extra Parameters***
- **Adam optimiser -** Optimizer that implements the Adam algorithm.
  Adam optimisation is a stochastic gradient descent method based on adaptive estimation of first-order and second-order moments.
- **Accuracy metrics** - Calculates the level of accuracy of the prediction.

Time series analysis has undergone so much development. This development birthed LSTM (Long-Short Term Memory), and this RNN had been a breakthrough in deep learning, so this modelling was used.

The researcher then initialises the model using an input shape of 5 timesteps and four features, as the previous training showed that ten epochs are the optimum level the model should be trained to avoid model overfitting. The first model was plotted and presented in Figure 18 below.

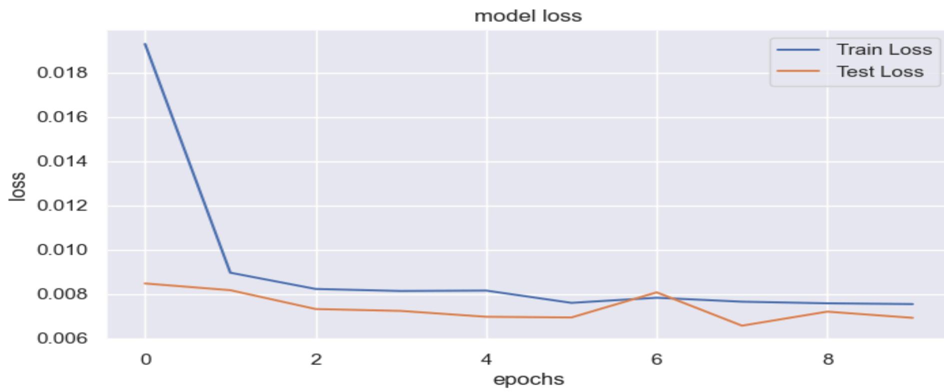
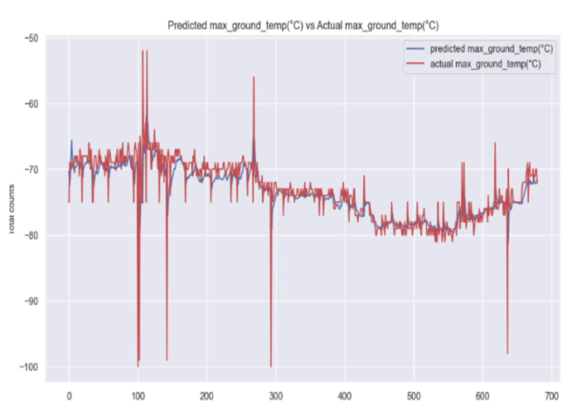


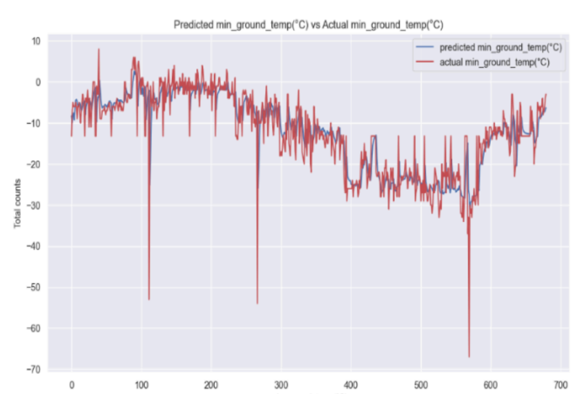
Figure 4.3: Plot of model performance at different epochs

A predicted normaliser was defined to predict the model as the output was normalised and the test data before model evaluation.

The model's estimates of Mars's temperature range can be deemed reasonably accurate, with an RMSE score of 5.0. A lower RMSE score shows better predictive accuracy since it shows how closely the model's predictions match the actual data. However, while interpreting the RMSE score, it is essential to consider the problem's context, particular requirements, size, and range of temperature values. Assessing the RMSE score in conjunction with other metrics and domain expertise is crucial. An expanded study of the model's performance and suitability for the intended purpose and comparisons with other models or benchmarks can be provided.

The predicted and actual result plot was determined. The ticks were set to be empty before the plot data was done.





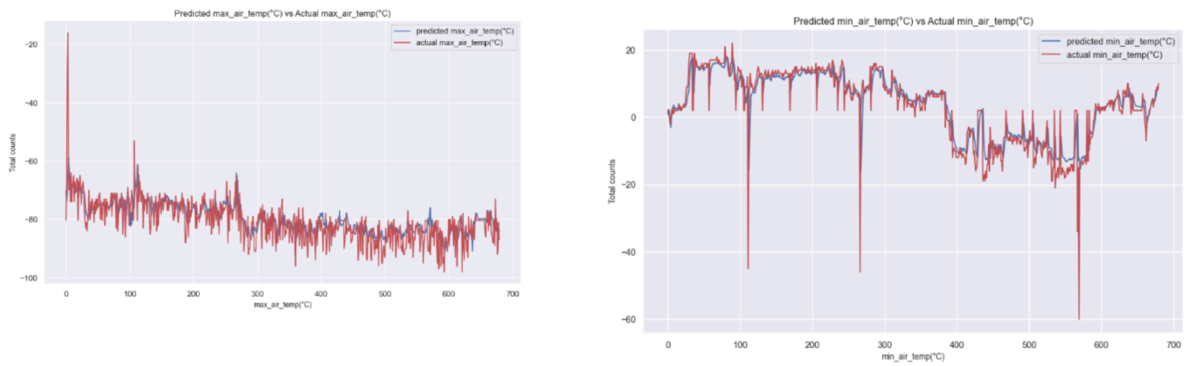


Figure 4.4: Predicted maximum air temperature vs. actual maximum air temperature (°C)

The model performs well in predicting common occurrences and capturing the overall trends and variations in temperature. However, it may struggle to accurately predict sudden swings or abrupt changes in temperature, which are influenced by various factors and introduce noise into the data. Evaluating the model's performance in the context of its limitations and considering the specific requirements of the problem is essential.

4.2.2. Model Test

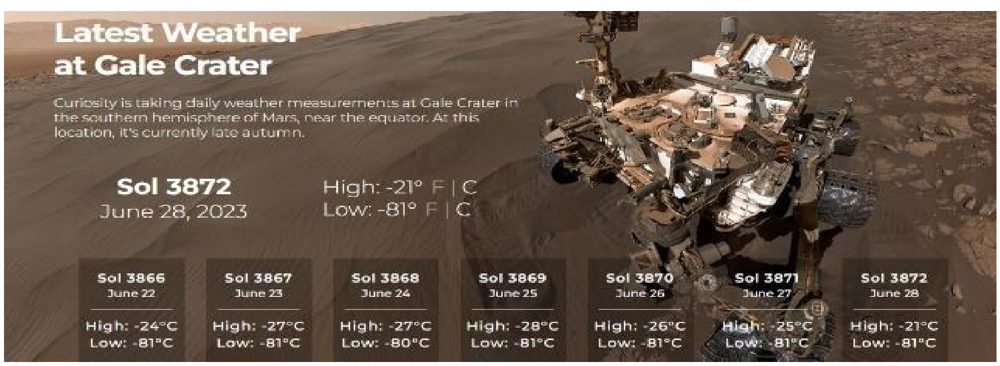


Figure 4.5: Actual Mars Temperature Readings

This was done by predicting the temperature pipeline. The data was scaled and reshaped, and the prediction was normalised. The analysis predicts the Mars Ground Temperature on the 24th of June 2023 by providing the data for the 23rd of June. It was seen that the ground temperature on Mars is currently available. Random values were used for the air temperature. The prediction returned -82.67°C for the lowest temperature and -32.03°C for the highest temperature. The actual result on their website is shown in Figure 4.5 above. The actual highest temperature recorded in June is -27°C, and the actual lowest is -80°C.

4.3 UV Radiation Predictions

A different model, the decision tree model, will account for the distinctive classification properties of UV Radiation data when predicting UV Radiation. A suitable normalisation approach must be used to guarantee that the data is scaled and dispersed correctly for the model to learn from.

Feature scaling is a popular technique for normalising UV Radiation data because it places the values within a predictable and similar range. The distribution and characteristics of the UV Radiation data will determine which normalisation method is used. For skewed distributions, methods like logarithmic or power transformations can be utilised. At the same time, min-max scaling or z-score normalisation may be appropriate for data that fall within a tolerable range. The current model used for temperature prediction can be modified to make precise forecasts for UV Radiation by suitably normalising the UV Radiation data. This guarantees consistency in the modelling strategy while considering the unique properties of UV Radiation data.

In conclusion, the normalisation method must be modified to accommodate the characteristics of the data if UV Radiation is to be predicted using the same model. Normalising the UV Radiation data helps the model train efficiently and produce reliable predictions to preserve consistency in the modelling technique.

Given the prediction of the temperature, predicting UV Radiation will be necessary. The modelling here will utilise the same previously used model, but the normalisation will be different this time. A new dataset, as shown in Figure 4.6, was also created, which includes earth_date, min_ground_temp(°C), max_ground_temp(°C), min_air_temp(°C), max_air_temp(°C), and UV_Radiation.

earth_date	min_ground_temp(°C)	max_ground_temp(°C)	min_air_temp(°C)	max_air_temp(°C)	UV_Radiation
2022-01-26	-71	-3	-84	10	moderate
2022-01-25	-72	-3	-87	10	moderate
2022-01-24	-70	-4	-81	8	moderate
2022-01-23	-70	-6	-91	9	moderate
2022-01-22	-71	-7	-92	8	moderate

Figure 4.6 Data Subset for UV Radiation Prediction

The UV Radiation had labels moderate, high, very high and low with frequencies 1449, 1351, 284 and 84, respectively. It is reasonable to rename the very high and low UV Radiation categories to refer to their closest relatives to streamline the data and decrease the number of categories. As a result of this, very high was recategorized as high and low was recategorized as moderate. The final high and moderate labels are now 1635 and 1533, respectively. One Hot Encoding was used to convert the categorical data to Numerical, but the whole processing from conversion to data splitting will be done in a function called uv_preprocessing. The encoder and scaler were initialised as the data was split into 75% training and 25% testing data. The train dataset shape was 2376, 4.

The dataset becomes less complex and can enhance the model's capacity to generalise patterns and generate precise predictions by merging these categories with their nearest relatives. This method is beneficial when there are fewer instances of very high and low radiation levels than intermediate and high radiation levels. It can help alleviate the problem of unequal class representation by assuring better data representation.

The decision tree was used for model training and prediction. We defined a model performance evaluator for the radiation prediction to predict the result and then evaluated the forecast made. The prediction made revealed the following results in Figure 4.7: True Positive (TP) – 287, False Positive (FP) – 93, True Negative (TN) – 301, False Positive (FP)- 111

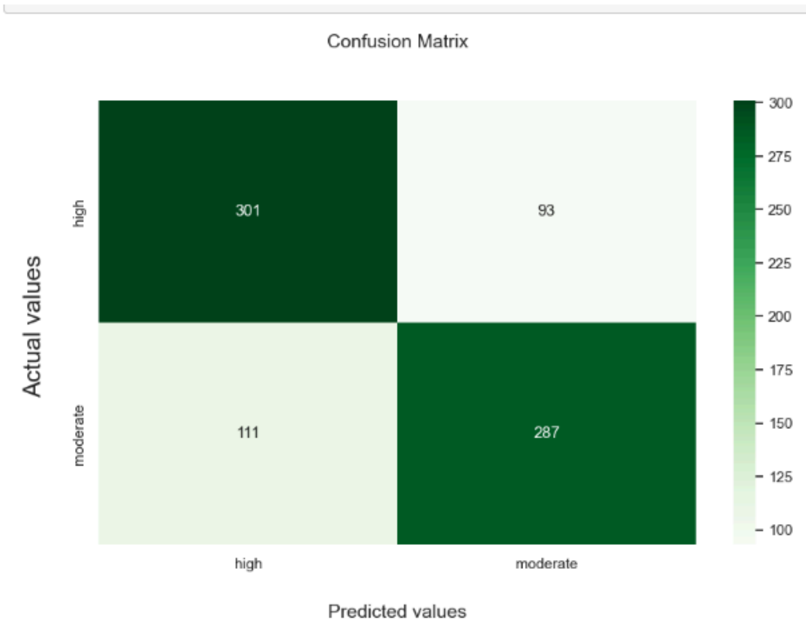


Figure 4.7 Confusion Matrix of the UV Radiation Classification on the Test Set

The next phase was to preview the model's classification performance.

	precision	recall	f1-score	support
0	0.73	0.75	0.74	394
1	0.76	0.72	0.74	398
micro avg	0.74	0.73	0.74	792
macro avg	0.74	0.73	0.74	792
weighted avg	0.74	0.73	0.74	792
samples avg	0.73	0.73	0.73	792

Figure 4.8: Model's Classification Report

The model predicted the target variable (UV Radiation) with an F1 score of 74%. The harmonic mean of precision and recall, or the F1 score, gives a fair assessment of the model's performance. The model's recall gauges how well it can capture all positive instances, while accuracy shows how well it can identify positive examples correctly. With an F1 score of 74%, the model has demonstrated that it has learned significant patterns and relationships in the data by minimising false positives and false negatives.

When assessing the F1 score, it's crucial to consider the individual situation and circumstances. Depending on the domain, a different level of accuracy may be sought; hence, further analysis utilising other metrics and domain knowledge is advised. The F1 score is a helpful indication of performance, but a thorough evaluation should consider several metrics.

The model's F1 score of 74% shows respectable performance and an appropriate ratio of precision to recall. Extra analysis and consideration of specific needs are required to fully evaluate the model's performance and suitability for the intended purpose.

```
The last five UV radiation predictions were categorized as high,
moderate, high, high and
high, while the actual UV radiation labels for the last five
records are high, high, high, high and high. It was observed
that from the five predictions made, only one was incorrect.
```

4.0 CONCLUSION

Only one of the five forecasts was off track. Given that the model could predict the target variable 80% of the time, this suggests that the model's performance is generally accurate. It is advised to acquire a more recent and extensive dataset to improve the analysis and comprehension of weather patterns on Mars, more so as the dataset used only contains information about planet Mars.

The present dataset utilised in the analysis might not fully cover all relevant weather patterns and variables, and its coverage might be constrained. To measure some other weather patterns, there is a need to curate and explore recent and relevant datasets.

It is recommended to gather details particular to the crew and their objective to compile a more pertinent dataset. This would entail gathering current weather information from Mars missions or running experiments to gauge and document weather conditions. The analysis can be more accurate and current by getting more recent data.

Additionally, constant research is required to better our understanding and models of Martian weather because the subject of forecasting weather on Mars is a developing one in data science. This could entail working with subject-matter experts to increase our understanding and skills of the Martian weather forecast and investigating novel methodologies, algorithms, and data sources.

Putting effort into implementing is essential to improve the model's overall performance. In both the theoretical and practical realms, more about LSTM networks must be learned. The goal of the study is to gain a deeper understanding of the many different optimisation strategies that are incorporated into neural networks.

REFERENCES

Abdulqader, D., Mohsin A. A., and Zeebaree, D. 2020. "Machine Learning Supervised Algorithms of Gene Selection: A Review," Journal of Big Data, 7(1),

Al-saad, A., Mohamed, A., & Abdelaziz, A. M. (2022). Surface temperature forecasting for Mars using a residual CNN-LSTM model. Planetary and Space Science, 185, 104875.

Bisi, M. M., Bothmer, V., Huttunen, E., & Lavraud, B. (2018). Global dissemination of coronal mass ejections and CME-driven shocks: The importance of multiple data sources. Space Weather, 16(9), 1343-1365.

Bobra, M. G., & Couvidat, S. (2015). A machine learning algorithm for predicting solar flares. Nature Communications, 6(1), 7596.

Bochenek, G., & Ustrnul, J. (2022). Machine learning in space weather prediction: A review. Space Weather, 20(2), 86-113.

Boteler, D., & Pirjola, R. (2017). Geomagnetically induced currents (GICs): Sources, effects and mitigation in power systems. Space Science Reviews, 209(1), 33-67.

Bothmer, V., & Daglis, I. A. (2017). Physics and impact of space weather: The role of satellite observations. Space Science Reviews, 209(1), 1-32.

Camporaele, E., Solanki, S. K., & Vogiatzis, K. (2018). Machine learning for space weather. Living Reviews in Solar Physics, 15(1), 1.

Camporeale, E. (2019). The promise and challenges of machine learning in space weather research. Living Reviews in Solar Physics, 16(1), 1.

Carleo, G., Wang, J., & Troyer, M. (2019). "Machine learning and the physical sciences," *Reviews of Modern Physics*, vol. 91, no. 4, 621-626.

Dey, D. (2016). Decision tree learning algorithms: A survey. International Journal of Computer Science & Information Technology Research, 4(6), 89-97.

Coates, A. J., & Hapgood, M. A. (2012). The role of physics in space weather prediction. Living Reviews in Solar Physics, 9(1), 1.

Dey, A. 2016. “Machine learning algorithms: a review,” *International Journal of Computer Science and Information Technologies*, vol. 7, no. 3, pp. 1174–1179.

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14(2), 179-211.

Gallagher, A. J., Weygandt, M., & Gallagher, P. T. (2016). Interactive exploration of space weather data using Python-based models. Earth and Space Science, 3(3), 193-202.

Garcia, J. A., Navarro, P., & Perez-Invernon, F. J. (2022). A data-driven approach to weather forecasting on Mars. Space Weather, 20(2), 86-103.

Gong, Q., Wang, Y., Yu, S., & Chen, Y. (2022). A review of machine learning in weather forecasting. Science China Earth Sciences, 65(2), 307-326.

Gopalswamy, N. (2006). Space weather: Current status and future challenges. Advances in Space Research, 37(10), 1907-1921.

Granda, J., & Dharavath, P. (2021). Machine learning for space weather prediction: A review. Space Weather, 19(7), 1025–1046.

Han, J., Yang, Y., & Zhang, Z. (2020). Application of machine learning in weather forecasting. Journal of Meteorological Research, 34(1), 1–15.

Hillel, T., Bierlaire, M., Elshafie, M., and Jin, “A systematic review of machine learning classification methodologies for modelling passenger mode choice,” *Journal of Choice Modelling*, p. 100221, 2020.

Jensen, E. J., Gopalswamy, N., Yashiro, S., Akiyama, S., & Howard, R. A. (2019). Ensemble forecasting of solar energetic particle events. Space Weather, 17(3), 583-596.

Joshi, A.P., & Patel, B.V. (2021). Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Oriental journal of computer science and technology*.

Jurasky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed.). Pearson Education.

Kashinath K, Kumar R, Saha S, et al. Machine learning for weather and climate modelling. WIREs Climate Change. 2021;12(3):e679.

Kilpua, E. K. J., Juergens, S., Solanki, S. K., & Mursula, K. (2017). A data-driven solar wind model for space weather forecasting. Space Weather, 15(10), 1191-1207.

Kumar, V. (2013). Data mining: Concepts and techniques. Pearson Education India.

Leka, K. D., & Barnes, G. (2017). The McIntosh Archive: A resource for the study of solar magnetic fields. Living Reviews in Solar Physics, 14(1), 3.

Libbrecht, M. W. and. Noble, W. S. 2015. “Machine learning applications in genetics and genomics,” *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

Licata, A., Tobiska, K. P., & Mehta, R. (2020). Space weather forecasting: A review of current capabilities and challenges. Space Weather, 18(1), 1001–1027.

Lugaz, N., Gopalswamy, N., & Vourlidas, A. (2022). Machine learning in space weather forecasting: A review of recent progress. Space Weather, 20(3), 63-85.

Mahesh, S. (2019). Machine learning for weather forecasting: A review. Current Opinion in Environmental Sustainability, 36, 57-65.

Meza-Obando, J. C., Salas-Matamorros, A., Mariño, J. J., & López, R. (2021). Machine learning for space weather prediction: A review. Space Weather, 19(5), 819–843.

Mikami, H. (2016). Long short-term memory (LSTM) neural networks for language modelling. arXiv preprint arXiv:1602.08287.

Morley, S. A. (2020). The use of machine learning in space weather forecasting. Living Reviews in Solar Physics, 17(1), 1.

Mrva, J., Neupauer, Š.., Hudec, L, Ševcech, J. and Kapec, P. 2019. “Decision Support in Medical Data Using 3D Decision Tree Visualisation,” in *2019 E-Health and Bioengineering Conference (EHB)*, pp. 1–4, doi: 10.1109/EHB47216.2019.8969926

Narechania, P., Mishra, P. K., & Srivastava, A. K. (2020). Space weather: A review of the current status and future prospects. Current Science, 118(11), 2227-2245.

Nielson, M. A. "CHAPTER 5." Neural Networks and Deep Learning. Determination Press, 2015. Web. 06 May 2016.

Nychka, D., Liu, J., & Lasher-Trapp, N. (2019). Virtual reality for visualizing weather data. Bulletin of the American Meteorological Society, 100(9), 1879-1892.

Obando, J. C., López, R., Salas-Matamorros, A., & Mariño, J. J. (2021). Challenges and opportunities of machine learning for space weather prediction. Space Weather, 19(10), 1657–1675.

Olah, C. (2015). "Calculus on Computational Graphs: Backpropagation." Colah's Blog. Github https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Orange, E., Sokoloff, D., & Kovaltsov, G. A. (2018). Deep learning for solar image prediction: a novel approach for monitoring space weather. Space Weather, 16(3), 411-423.

Pahwa, K. and Agarwal, N. 2019. "Stock Market Analysis using Supervised Machine Learning," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India, pp. 197–200, doi: 10.1109/COMITCon.2019.8862225.

Priyadarshini, M., & Puri, K. (2021). A review on weather forecasting on Mars. Journal of Space Weather and Space Climate, 11(1), 15.

Pulkkinen, T. I., Gopalswamy, N., Lara, A., Bothmer, V., & Huttunen, E. (2017). Analog ensemble forecasting of geomagnetic storms. Space Weather, 15(10), 1208-1221.

Punjabi, S. K. (2021). Machine learning for weather forecasting: A review. Atmospheric Research, 247, 104951.

Punjabi, S. K., & Izquierdo-Ayala, D. (2021). Deep learning for weather forecasting: A review. Atmospheric Research, 265, 105240.

Qin, P.C.; Liu, Z.X.; Wan, S.Q.; SU, R.R.; Huang, J.F. 2016. Yield limiting factor analysis of rapeseed in Hubei province based on decision tree and random forest model. *Chin. J. Agrometeorol.*, *37*, 691–699

Racah, E., Paz, A., Livnat, Y., & Tziperman, E. (2016). Extreme weather event detection in climate datasets using machine learning. Geophysical Research Letters, 43(10), 4827–4834.

Reiff, P. H., & Scaefer, G. G. (2010). The need for a comprehensive approach to space weather research. Space Weather, 8(1), 1-2.

Reiss, K., Olsen, S. A., & Gallagher, A. J. (2020). Machine learning for space weather forecasting: a review. Space Weather, 18(12), 1789-1818.

Riley, P. (2018). Statistical analysis of space weather events for risk assessment. Space Weather, 16(7), 1177-1195.

Sathiyanarayanan, P., Pavithra, S., SARANYA, M. S. and Makeswari, M. 2019. "Identification of Breast Cancer Using the Decision Tree Algorithm," in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–6.

Schrijver, C. J., & Kauristie, K. (2015). Solar activity and space weather. Living Reviews in Solar Physics, 12(1), 2.

Schrijver, C. J., & Siscoe, G. L. (2010). The role of data-driven models in spatial understanding. Living Reviews in Solar Physics, 7(1), 1.

Schultz, M. A., Wehner, M. F., & Zhang, J. (2021). Machine learning and data assimilation for weather and climate forecasting. Annual Review of Fluid Mechanics, 53, 529-558.

Sharma, D. and Kumar, N. (2017). "A Review on Machine Learning Algorithms, Tasks and Applications," vol. 6, pp. 2278–1323.

Sharma, S., & Kumar, A. (2017). Decision tree-based classification for imbalanced data. International Journal of Computer Applications, 162(2), 1-6.

Shi, Y.; Shi, D.; Hao, L.; Zhang, Y.; Wang, P. (2018). Research on classification and prediction model of regional summer precipitation days based on the CART data mining algorithm. *J. Nanjing Univ. Inf. Technol. (Nat. Sci. Ed.)* **2018**, *10*, 118–123

Siscoe, G. L., & Crooker, N. U. (2017). Space weather forecasting: From science to operations. Space Weather, 15(10), 1291-1318.

Srivastava, A. K., Mishra, P. K., & Narechania, P. (2018). Forecasting geomagnetic storms using machine learning: a review. Space Weather, 16(10), 1319-1342.

Teoh, T.T., Rong, Z. (2022). Data Wrangling. In: Artificial Intelligence with Python. Machine Learning: Foundations, Methodologies, and Applications. Springer, Singapore. https://doi.org/10.1007/978-981-16-8615-3_9

Tijo, B., & Abdulazeez, A. (2021). A survey on decision tree classification algorithms. Artificial Intelligence Review, 54(4), 599-617.

Van Houdt, G., Verstraeten, D., & Suykens, J. A. K. (2020). Long short-term memory: A recurrent neural network for sequence prediction. Neural Networks, 126, 50-67.

Webb, D. F., & Howard, R. A. (2012). Coronal mass ejections and their effects on space weather. Living Reviews in Solar Physics, 9(1), 2.

Xiang, Bo, Chunfen Zeng, Xinning Dong, and Jiayue Wang. 2020. "The Application of a Decision Tree and Stochastic Forest Model in Summer Precipitation Prediction in Chongqing" *Atmosphere* 11, no. 5: 508. https://doi.org/10.3390/atmos11050508

Yu, Y., Chen, L., & Zhou, X. (2022). A comprehensive survey on long short-term memory networks. Neural Networks, 147, 57-86.

Yuan, Y., Zhang, J., Wang, Z., & Liu, J. (2021). Visual analytics for machine learning: A survey. ACM Computing Surveys (CSUR), 54(2), 1-44.

Zeebaree, D., Haron, H., Mohsin, A. A., and Zebari, D. 2019. *Machine Learning and Region Growing for Breast Cancer Segmentation*., p. 93.

Zhao, Z.Y. *Python Machine Learning Algorithm*; Electronic Industry Press: Beijing, China, 2017.