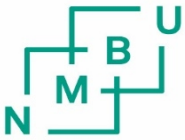



Norges miljø- og
biovitenskapelige
universitet

**Master's Thesis 2024 30 ECTS**
Faculty of Science and Technology

# Investigating the Viability of Machine Learning for the Prediction of Icing Occurrences at Airports

Tonje Martine Lorgen Kirkholt
Data Science

# Acknowledgements

This thesis would not have been possible without the guidance and help of several individuals, each contributing in their own way by sharing their knowledge, care, and support.

I would like to express my deepest appreciation to my supervisor, Associate Professor Dr. Eirik Valseth, for his invaluable patience and feedback on this thesis. Thank you for the reassurance you gave through encouragement and guidance whenever it was needed and for sharing your time and knowledge.

I would like to extend a big thank you to The Norwegian Meteorological Institute for inspiring me with the topic of this thesis and providing me with access to the necessary data and extra computational power. I want to show my gratitude to Dr. Jean Rabault for his engagement, thorough guidance, and motivation. A huge thank you for your patience, for encouraging weekly meetings, and for providing so much of your time for the guidance provided. A special thanks to John Bjørnar Bremnes for sharing your knowledge in machine learning and providing insights into similar work as a research scientist at MET. I am also thankful for the rest of the team and co-workers at MET ITGEO, for all the support and positivity.

I am forever grateful for my friends and SO, who never stop encouraging, supporting, and cheering. Thank you for believing in me.

The most significant acknowledgment is reserved for my parents; the constant drive and motivation I find in myself would not have been the same without the support you have provided throughout the years. I am forever in debt. Thank you.

**Abstract**

Ensuring the functionality of airport operations amidst changing weather conditions is crucial for maintaining operational efficiency and ensuring safety. This thesis investigates using Machine Learning (ML) techniques to predict icing weather events at airports, aiming to enhance aviation safety through improved forecasting. Leveraging input data from Numerical Weather Prediction (NWP) models and real-time observations extracted from Meteorological Aerodrome Reports (METAR), probabilistic classifiers are developed and further assessed through their efficacy in providing reliable predictions for freezing weather occurrences. Due to the rarity of these events, the resulting dataset is inherently imbalanced, necessitating heavy downsampling to facilitate efficient model training. Despite the small dataset size, the developed models demonstrate promising capabilities, exhibiting notable improvements in reliability and accuracy, particularly within the temporal models. Notably, even with minimal training data, the models accurately predict freezing weather occurrences up to five timesteps ahead, each representing one hour.

Hence, this study emphasizes the importance of collaboration between ML experts and domain specialists in aviation meteorology to gain deeper insights and refine the models. The results serve as a solid foundation for reflection and offer valuable suggestions for future research directions to enhance ML models' predictive capabilities in this domain.

# Contents

# 1 Introduction

The collective occurrence of daily atmospheric events dictates the prevailing weather conditions. Forecasting the weather is essential for ensuring safe flight operations by providing pilots with the necessary details to make informed decisions throughout their flights. The warming climate has increased hazardous weather events of different types and intensities [9]. These events encompass heavy rainfall, tornadoes, thunderstorms, and freezing weather, each carrying potential risks such as cancellations and delays [19]. In severe situations, extreme weather events can cause damage to aviation systems, aerodromes, infrastructure, and human life.

Over the last few years, an increase in the intensity of the events and a shortened time gap between occurrences has been observed [12]. The concept of weather whiplash, a sudden and quick change in types of extreme weather, presents challenges in producing accurate and long-lasting forecasts [9]. Rapid temperature fluctuations, swinging above and below freezing temperatures, can increase the risk of unprepared freezing weather incidents.

To highlight the importance of better understanding and further improving prediction of extreme weather events, Sillmann et al. [44] outline five key questions addressing scientific challenges in predicting weather extremes:

a) *What are relevant definitions of extremes on the respective time scales?*

b) *What are the necessary observations and model output requirements to analyze these extremes?*

c) *What are the processes driving these extremes and their changes?*

d) *How do we best evaluate these extremes (including relevant processes)? (i.e., is the model right for the right reason)*

e) *What are relevant sources for predictability of these events that can support the attribution, prediction and projection of these extremes?*

Although numerous studies demonstrate that ML can enhance the prediction of typical weather events, there is still a need for more research on extreme weather events [47]. The scarcity of samples containing extreme weather incidents poses challenges, potentially causing ML models to struggle or fail in the worst case. Despite this, integrating deep learning with meteorological science holds promise and has already contributed to better extreme weather forecasting [16].

## 1.1 Background

The information in an aviation weather forecast differs from the standard forecasts that most people can read and understand, as these are encrypted strings of text consisting of letters and numbers. There are several types of weather reports, but the most common ones are Meteorological Aerodrome Reports (METAR), Terminal Area Forecast (TAF), and Significant Meteorological Information (SIGMET). Each serves its purpose, where the main difference is the period in which the report describes the weather. METAR contains the weather status at a specific time, issued once every half hour. TAFs give us an overview of how the weather will develop over the next 24 hours, some up to 30. When discussing SIGMETs, these reports warn about upcoming extreme weather events in a given area in flight.

The quality and information communicated through these messages are crucial and highly dependent on weather observations being as accurate as possible. The Convention on International Civil Aviation (ICAO) issues guidelines for constructing METAR and TAFS in Annex 3, titled "Meteorological Service for International Air Navigation" [24].

In a TAF or METAR message, freezing weather is indicated by abbreviations like FZRA (freezing rain), FZFG (freezing fog), and FZDZ (freezing drizzle). Freezing weather indicates that the precipitation is supercooled, meaning the temperature is below freezing, but the water droplets have not solidified. It freezes in contact with the ground or other surfaces with temperatures lower than 0°C. Situations where icing occurs on an aircraft, can cause a buildup of weight and drag, further reducing the lift for takeoff [20] and in worst cases lead to catastrophic disasters [7]. In Figure 1, we observe a simple drawing outlining this process.

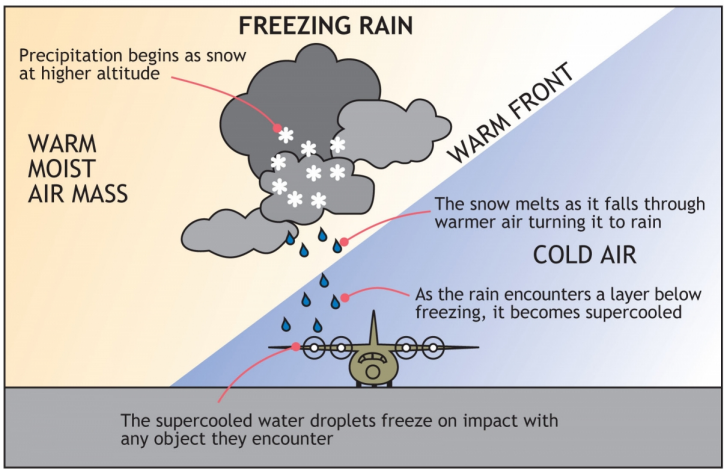



Figure 1: Visual description of how freezing rain occurs. Reproduced from `https://skybrary.aero/articles/freezing-rain`, accessed on 14 December 2023.

Utilizing advanced technology enhances our ability to predict extreme weather with increased accuracy [28]. A thorough examination of freezing weather events is critical for improving preparedness and mitigating risks, contributing to the safety of aviation systems in the face of challenging weather conditions.

## 1.2 Contextualizing the Objective

Today, traditional forecast methods are done by observing the current condition of the atmosphere and further predicting future atmospheric state. The NWP models process the current weather observations to forecast future weather by running large-scale simulations that describe how the atmosphere moves and changes. From these results, the goal is to represent the current and future atmospheric phenomena as accurately as possible.

Meteorologists rely on these values to issue METAR reports that precisely describe the current weather conditions. This thesis explores the feasibility of employing ML to predict freezing weather occurrences at airports. With this in mind, the central question arises: Can these parameters serve as inputs for an ML model and further be trained to generate reliable probabilities for freezing weather occurrences at airports? A reliable prediction would entail a probability that aligns closely with the observed occurrence rate.

As a potential tool for meteorologists, this could enhance the accuracy of METAR reports and facilitate more informed decision-making processes.

To address this inquiry, this thesis undertakes and presents a comprehensive process involving data extraction, preprocessing, and training of ML models, serving as a foundation for further evaluation of this objective.

## 1.3 Limitations of the Study

### 1.3.1 Skewed Class Proportions

The abbreviation FZ in METAR and TAF reports, indicating freezing conditions, is not frequently encountered. Consequently, it was impossible to avoid significant class imbalance when creating the dataset with a column indicating the presence or absence (True or False) of FZ occurrences in the extracted METAR messages. The infrequent occurrence of FZ represented as the boolean value True in the mentioned column led to skewed class proportions, further impacting the effectiveness of Neural Networks (NN) trained on this dataset for predicting these exact occurrences.

Initially, training the NN using the original imbalanced dataset led to an apparent bias toward predicting the majority class, overlooking instances of the minority class. To address this issue, downsampling the majority class was implemented. This approach ensured a more balanced representation of classes but significantly reduced the size of the training dataset.

The test dataset obtained after splitting the dataset remained unchanged, and this part of the process was completed using imbalanced data. While downsampling improved the model's performance, extracting data from a broader period or upsampling the data could further enhance the model's performance. This would enlarge the size of the minority class, ensuring a higher number of occurrences of FZ and further provide the model with more data to learn from, thereby potentially improving its ability to accurately predict the likelihood of rare events.

### 1.3.2 Auto Generated METAR Reports

In the dataset, we classify METAR reports into four types based on their generation method and whether they invalidate and correct a previous observation. Notably, several airports utilize automated software to generate METAR reports, meaning that no human intervention occurs in the making

of the report. However, according to meteorologists at MET, many of these reports, mainly those forecasting freezing weather, often contain inaccuracies due to the model's limitations in effectively capturing such conditions. The airports responsible for most reports forecasting freezing weather predominantly rely on auto-generated messages and, consequently, make up many of the registered cases for the occurrence of FZ in a METAR.

During model training, efforts were made to remove all the "AUTO" metartype instances, resulting in a separate dataset. This action significantly reduced the already small minority class by almost half the observations. Subsequent training and testing of the model with further reduced data revealed that the model's performance had no immediate drastic or negative change. As a result, this dataset was incorporated into the processes involving the persistence and instantaneous models. However, considering the time constraints, the main focus was directed towards the two datasets where these instances were retained for the temporal models.

## 1.4 Structure of the Thesis

To enhance the objective's theoretical foundation, Sections 2.1 and 2.2 delve into weather forecasting for aviation and the underlying causes of freezing weather. Section 2.4 provides a brief overview of machine learning concepts, paving the way for a detailed exploration of Artificial Neural Networks and their components in Section 2.5.1 The critical field of model evaluation is assessed in Section 2.6, which introduces techniques such as confusion matrices, ROC, and calibration curves in combination with AUC and AUC-PR scores. Section 3 shifts the focus to the practical aspects of the research. It begins with an examination of the processes involved in data acquisition and preparation of the extracted METAR reports and meteorological parameters in Section 3.1. Sections 3.2 and 3.3 elaborate on the development process and decision-making in constructing the probabilistic classifier.

Then, there is a transition to Section 4, which thoroughly presents the results obtained from the trained and tested models. This section comprehensively compares the persistence, instantaneous, and temporal predictions, with a detailed evaluation and discussion of these results in Section 5. Finally, Section 66 encapsulates the thesis, offering conclusions and reflections illuminating potential approaches for future work.

# 2 Theory

This chapter will provide the necessary theory for understanding the concepts behind weather forecasting for aviation and the different methods used today. It will also present the principles behind NWP prediction. Last, we dive into an explanation of ML and Artificial Neural Networks (ANN) as the primary tools for the prediction work in this study.

## 2.1 Weather Forecast for Aviation

Annex 3 of the Convention of International Civil Aviation (ICAO) outlines guidelines for international standards and recommended practices regarding weather forecasting for aviation [24]. According to these specifications, a weather forecast comprises a statement detailing expected conditions for a specified time, area, or portion of airspace.

The following three sections describe the key messages and reports that aviation weather forecasting relies on, including METAR, TAF, and SIGMET. The importance of this information is underscored by Sections 2.1.1 and 2.1.2 of ICAO Annex 3 [24], which articulate its objective: to provide essential meteorological data to various stakeholders involved in international air navigation:

*"2.1.1 The objective of meteorological service for international air navigation shall be to contribute towards the safety, regularity, and efficiency of international air navigation."*

*"2.1.2 This objective shall be achieved by supplying the following users: operators, flight crew members, air traffic services units, search and rescue services units, airport managements and others concerned with the conduct or development of international air navigation, with the meteorological information necessary for the performance of their respective functions."*

### 2.1.1 Meteorological Terminal Air Report (METAR)

A Meteorological Terminal Air Report (METAR) is a short report that gives information about the current weather conditions at a specific location, where usage is intended mainly for pilots and meteorologists. It is issued every half or whole hour and is valid until the next issued METAR. The forecast con-

tains a lot of information compressed into a short text string of coded data. Table 1 shows the specific parts of a METAR issued in ICAOs Annex 3 [24]. Figure 2 illustrates a sample METAR message, deciphering its components and corresponding forecasts.

Table 1: Overview of the content of a METAR message

| **Content** | **Description** |
|---|---|
| *Type of report* | Informs about specific report type. Possible reports are METAR or SPECI. |
| *Location Indicator* | A unique code used to identify the station for which the report is forecasted for. |
| *Time of the observation* | Two first digits are the date and the four last represent the time. |
| *Identifier if modified report* | Additional information about the report, either informing about an automated (AUTO) or a corrected (COR) report. |
| *Wind* | Surface wind direction and speed. |
| *Runway isibility* | Given in metres. |
| *Present Weather* | Present weather occurring at the specific aerodrome. Describing, if present, (freezing) precipitation with intensity, thunderstorms, fog, or and freezing fog. |
| *Sky condition* | Description of cloud cover and base of clouds in hundreds of feet. |
| *Temperature and dew point* | Air temperature and dew point temperature in degrees Celsius, with temperature first. |
| *Pressure* | Current pressure at mean sea level (QNH). Computed and reported in hectopascals. |
| *Remarks* | Additional remarks about significant weather phenomena not included in other sections. |


Figure 2: An example of a METAR message and how to decode it.

### 2.1.2 Terminal Aerodrome Forecast (TAF)

Terminal Aerodrome Forecasts (TAF) are concise messages that give information about the expected meteorological conditions for the surrounding area of a specific airport. The TAF forecasts have a validity period of 24-30 hours and cover a radius of approximately 8 kilometers (5 miles) around the specified airport. Typically, a TAF is updated four times daily: 00, 06, 12, and 18. The message contains the same information as the METAR report, details found in [24].

### 2.1.3 Significant Meteorological Information (SIGMET)

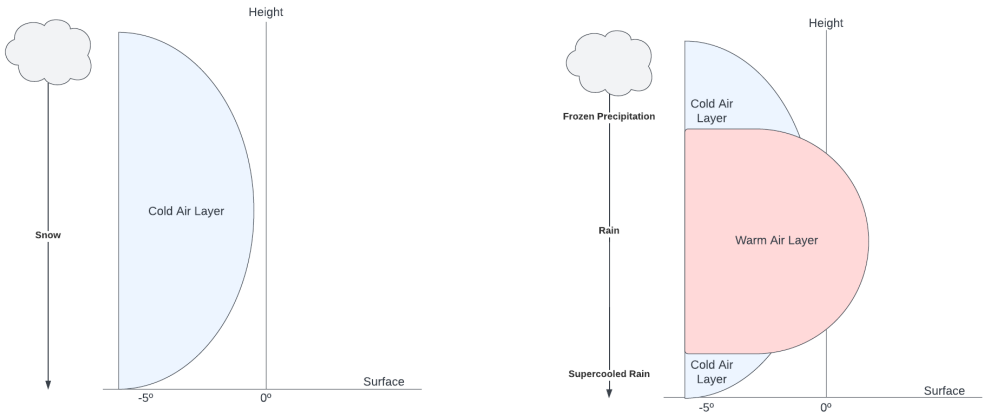
A Significant Meteorological Information (SIGMET) report is issued to inform aircraft operators about prevailing or anticipated weather phenomena along a flight route that could affect the safety of aircraft operations [24]. These reports can remain valid for up to four hours and are canceled when the specific weather phenomena cease or are no longer expected. However, in certain circumstances involving volcanic ash clouds or tropical cyclones, the report's validity may be extended by two hours.

## 2.2 Freezing Precipitation

Freezing precipitation is a weather phenomenon in which frozen precipitation encounters atmospheric layers with divergent temperature profiles, where the air temperatures are between above- and below-freezing temperatures [6]. This results in the formation of icy coatings on the affected surfaces and can further present hazards to infrastructure.

When frozen precipitation, such as snow, encounters a layer of air above freezing temperature, it will melt and turn into rain or drizzle. As this liquid precipitation descends, it encounters a sub-freezing layer of air near the ground. Here, the temperature is low enough to cause the liquid to become supercooled, where the liquid remains unfrozen despite being below the freezing threshold [18]. Upon contact with objects or surfaces with temperatures below freezing, the supercooled droplets will instantly freeze and form a layer of thin ice. This freezing process can also occur with fog. When fog forms in sub-freezing conditions, its liquid water droplets remain liquid. Upon contact with surfaces that have a temperature below 0 degrees, it freezes and creates a layer of ice crystals.

The common denominator for characterizing freezing weather is the process by which freezing precipitation melts into liquid precipitation before turning into ice again. Figures 3a and 3b presents a rough sketch of how the temperature profiles for the occurrence of snow and freezing rain can be vizualised.




(a) Temperature profile for snow

(b) Temperature profile for freezing rain

Figure 3: Visual comparison of temperature profiles for the occurrence of snow versus freezing rain. Based on Figure 2, Forbes et al. [18]

### 2.2.1 Freezing Precipitation Effects on Airports and Aircraft

Weather and meteorological conditions present significant challenges for aircraft in the air and ground. Unfavorable weather conditions can interfere

with safe takeoffs and landings, disrupt flight paths, cause delays, and affect overall flight operations. Therefore, preparing for existing weather conditions and anticipating potential changes is crucial for ensuring flight safety. Pilots rely highly on issued weather reports such as METAR, TAF, and SIGMET, presentes in Sections 2.1.1, 2.1.2, and 2.1.3.

Freezing weather, such as freezing rain, drizzle, and fog, can cause icing on the aircraft and the runway, affecting the aircraft in several ways. If ice builds up on the aircraft, it increases its overall weight, making it harder to take off because of the increased drag and decreased lift. Ice can also affect the propellers by coating the blades, reducing the overall thrust. A growing coat of ice on the wings and tail can disrupt proper airflow. There are different severities of icing on an aircraft, where moderate, heavy, and severe icing clearly affects the aircraft's performance [8].

The most severe consequences of aircraft icing have resulted in tragic accidents, with several documented cases. One such incident occurred in Canada on December 13, 2017, involving a Saskatchewan airline [10]. The aircraft, compromised by ice buildup, experienced a loss in altitude shortly after takeoff, leading to a crash that tragically resulted in the death of a passenger. Similarly, on January 9, 2011, in Iran, a Boeing 727-200 crashed due to severe icing, which led to a blockage of air into the engines and a subsequent loss of thrust [45]. As a result, the aircraft lost power and began to descend, ultimately crashing and claiming the lives of eight out of the nine crew members and 70 out of the 96 passengers aboard.

## 2.3 Numerical Weather Prediction

Numerical Weather Prediction (NWP) is a central component for predicting the weather based on current weather conditions. NWP is based on the numerical solution of the partial differential equations that govern the behavior of the atmosphere [37]. NWP models compute future values of vital atmospheric parameters by utilizing initial values derived from meteorological observations. Integration of these models into weather forecasting has significantly improved prediction accuracy, equipping meteorologists with advanced tools for analysis and forecasting [13].

### 2.3.1 MetCoOp Ensemble Prediction System (MEPS)

In 2010, MET Norway and the Swedish Meteorological and Hydrological Institute (SMHI) initiated a collaboration called MetCoOp, short for Meteorological Cooperation on Operational Numeric Weather Prediction [31]. Following the operationalization of the MetCoOp Ensemble Prediction System (MEPS) by the end of 2016, the collaboration expanded with the inclusion of the Finnish Meteorological Institute (FMI), which joined in 2017. In 2019, the Estonian Environment Agency (ESTEA) joined the cooperation [23]. The project aims to expand further, with plans to involve several other countries.

The MEPS files utilized for data extraction in this thesis comprise 30 lagged ensemble members within a 6-hour time window, operating at a resolution of 2.5 km [3]. For this thesis, emphasis is placed on the control member within the ensemble, designated as member 0, which serves as the primary source for data extraction. This control member is derived from the most accurate analysis and represents the most probable forecast.

Model runs are completed, and new forecasts are generated based on observations every six hours, precisely at 00, 06, 12, and 18.

## 2.4 Machine Learning

Machine Learning (ML) springs out from Artificial Intelligence (AI) and computer science, focusing on using data and algorithms to mimic how human beings process and learn new information.

### 2.4.1 Learning Strategies and Methods

In the domain of ML, we have a selection of algorithms that can be employed to tackle the specific tasks at hand, broadly categorized into four primary types: supervised, unsupervised, semi-supervised, and reinforcement learning. However, for the scope of this discussion, the focus will primarily be on supervised learning, as it is the chosen algorithm for the specific methodologies utilized within the study presented in this thesis.

Supervised learning is a fundamental approach in ML, where the data used for training is labeled beforehand [38]. The model is trained to predict a specific label or target value associated with each data point. The primary objective of supervised learning is to train the model to recognize underlying

patterns in the input features and, with a high degree of precision as the aim, predict the associated labels on new, unseen data. Figure 4 depicts this process.

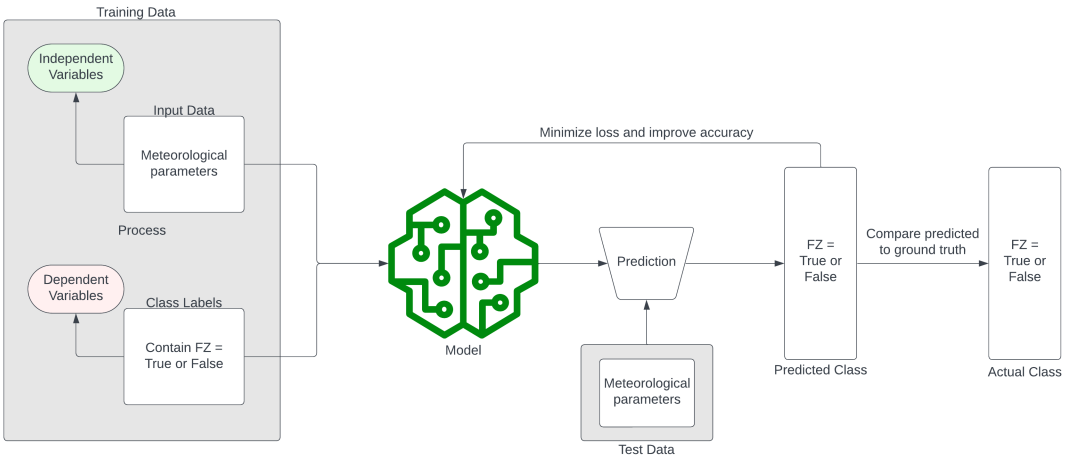



Figure 4: Visual description of the process of a supervised machine learning model.

## 2.5 Model Selection

In this thesis, the motivation behind the model selection is guided by an evaluation of the research objective, the desired type of outcome and prediction, and the characteristics of the data.

The objective of this thesis revolves around predicting the likelihood of icing weather incidents at airports based on observed meteorological parameters and historical METAR reports. The predicted variable will be based on boolean variables for the occurrence of FZ in a METAR report, making it a binary classification problem because we want to differentiate between True and False.

The decision to employ an Artificial Neural Network (ANN) is grounded in research that has demonstrated promising weather forecasting outcomes using this model implementation [34, 1, 30]. The ANN's ability to capture nonlinear relationships within the data is particularly appealing, presenting itself as a good fit for the outlined objective. While the objective of this study does not involve traditional weather forecasting, it utilizes meteorological

parameters. It further explores the relationship between them to predict the probability of a specific weather phenomenon.

### 2.5.1 Artificial Neural Network

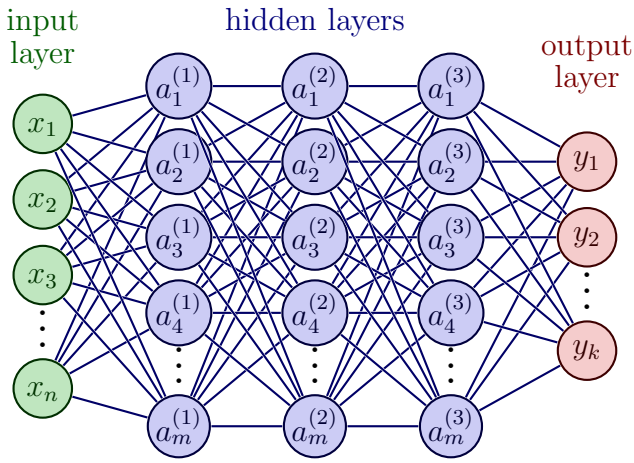



Figure 5: Fully connected neural network with three hidden layers.

A neural network has three main layers: one input, one or more hidden, and an output. The input layer introduces the information into the network, the hidden layers perform computations, and the output layer presents processed information.

Figure 5 presents a fully connected network, meaning that each neuron connects with every node in the preceding layer or output.

The input layer consists of nodes, $x_1, x_2, ..., x_n$, where $x_n$ is the input from the $n$-th feature or neuron in the input layer. In forward propagation, each unit in the layer $l$ is connected to all the units in layer $l+1$ through a weight coefficient. If we want to calculate the activation unit of the hidden layer $a_1^{(1)}$, we do the following:

$$z_j^{(l)} = \sum_{i=1}^{N^{(l-1)}} w_{ij}^{(l)} \cdot a_i^{(l-1)} + b_j^{(l)} \tag{1}$$

$$a_j^{(l)} = \phi(z_j^{(l)}) \tag{2}$$

Figure 6 shows what this looks like for the first layer in our neural network in Figure 5.

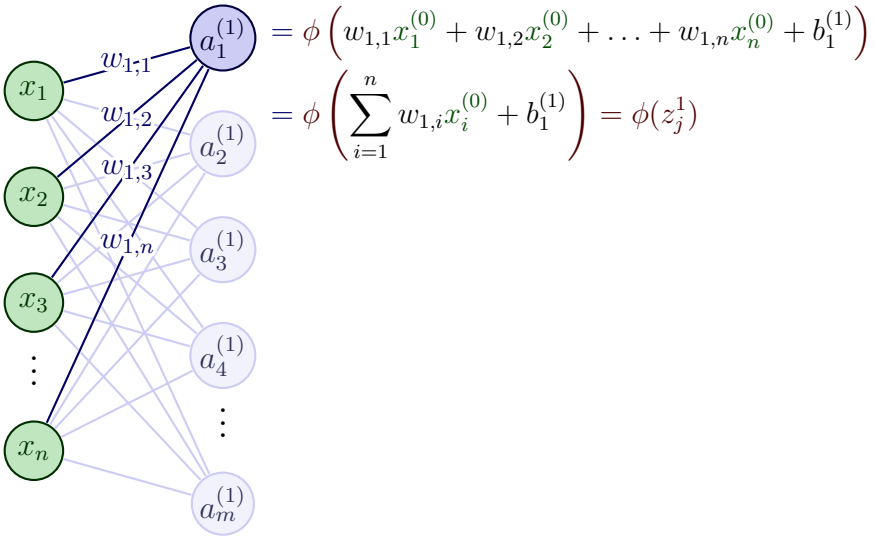



Figure 6: Explanation of activation function in neuron in addition to one full layer.

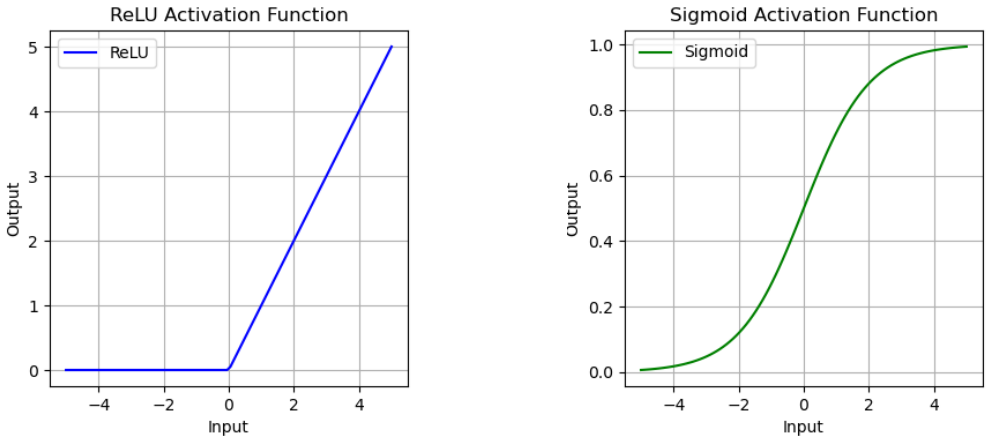
The activation function, $\phi(z)$, depends on what you want to predict. Incorporating an activation function introduces non-linearity into the output of each neuron, where the lack of an activation function would reduce the network to a simple linear regression model.

### 2.5.2 Activation Function

In a neural network, the activation function determines the output of a neuron based on the input. The activation functions compute the weighted sum of inputs and biases, acting like a decision-maker for whether or not the neuron should be triggered to send a signal to the next layer in the network. While an activation function can be linear, the real strength lies in the option to introduce non-linearity in a model [15]. Non-linearity allows the model to learn more intricate patterns and relationships within the data. Each activation function offers various properties, and selecting the most optimal one depends on the specific objective and dataset.

In the context of the final ANN presented and discussed in Sections 3.2 and 3.2.1 of this thesis, the focus lies on the chosen activation functions, which both introduce non-linearity to the model: the Rectified Linear Unit (ReLu) and the Sigmoid function. Figures 7a and 7b illustrate the ReLu and Sigmoid plots. In Equation (1) in section 2.5.1, the variable z represents

the weighted sum of inputs and biases. The specific activation function is computed based on this sum, as shown in Equation (2).




(a) $\phi(z) = \frac{1}{1+(e)^{-z}}$ (b) $\phi(z) = max(0, z)$

Figure 7: ReLu and Sigmoid activation functions.

The Sigmoid function is often used in the output layer of a neural network because of its range. Producing outputs between 0 and 1 can be interpreted as probabilities, which makes it a good match for binary classification and the objective of this thesis [43].

The ReLu is computationally efficient compared to other activation functions because of its simple mathematical structure, where the output value is always one for positive values and otherwise returned as zero. This property of ReLu addresses the vanishing gradient problem, where the partial derivative of the activation function diminishes to near-zero values, resulting in it vanishing and terminating weight updates [38]. In ReLu, the derivative remains constant at one for positive inputs, ensuring that the gradients do not vanish. Because of the output values produced by ReLu, it is primarily suitable as an activation function in the hidden layers [43, 38].

### 2.5.3 Regularization

Finding the right balance between complexity and performance in an ML Model can be challenging. Section 2.6.1.1 highlights some commonly encountered issues in this process, such as overfitting and underfitting. Regu-

larization is a technique used to tackle overfitting, which works by controlling the values of the model's parameters by introducing specific rules or constraints. This prevents the model from learning overly complex patterns, including noise and irrelevant information, ultimately improving its ability to generalize to new data.

Of the various regularization techniques available, this work focuses on L1 regularization, early stopping, and batch normalization as the primary methods.

The L1 regularization method is also known as LASSO, an abbreviation for "the least absolute shrinkage and selection operator." The penalty term added by LASSO is the absolute value of the weights multiplied with a regularization parameter that can be adjusted in strength to control its overall effect [38, 33]. LASSO also contributes to feature selection by shrinking the coefficients of irrelevant features toward zero, because large absolute values of the model's parameters are penalized.

Early stopping is another technique employed to mitigate overfitting. This feature monitors the model's performance during training and terminates the process when it no longer observes further improvement, overriding the pre-defined number of training epochs This intervention occurs if the performance improves on the training data but worsens on the validation data after an initial improvement period [36]. A stopping criterion determines the specific conditions under which training ceases.

Batch normalization is usually added between the hidden layers to accelerate and stabilize the training process [42]. It normalizes the inputs by calculating the mean and variance for each mini-batch, where a mini-batch is a smaller portion of the dataset used in one training iteration [25, 38].

### 2.5.4 Loss Function

During training, the main goal is to work towards the most efficient model. Part of this is comparing the predicted values with the actual ground truth and minimizing the difference between them.

Loss functions are chosen based on the objective. Since we are dealing with a classification problem with probabilistic outputs, binary cross-entropy is the loss function used in the final ANN model [38].

### 2.5.5 Optimization

Training the neural network involves optimizing the parameters appropriately to work towards a minimized loss function. Among the available options for optimization algorithms, Adaptive moment estimation (Adam) is a popular and widely used method [38]. Adam is a gradient-based optimization algorithm known for its computational efficiency and ease of tuning [26]. It only requires specifying a learning rate before training.

## 2.6 Model Evaluation

A comprehensive evaluation of predictive models, from training to actual testing, is essential for understanding their performance and generalization capabilities. Rigorous evaluation throughout the training process ensures that the model learns from data effectively. Further, estimating the performance when fed with new unseen data is equally important, providing insights into its predictive accuracy and ability to generalize new input.

The following two sections present the evaluation methods utilized for training and testing the model employed in this thesis.

### 2.6.1 Training Data

Training and testing a model by partitioning the data into separate datasets for training and testing is also commonly known as the holdout method [38]. This is one of the simplest methods to evaluate a model's performance by exposing it solely to a portion of the data during training and evaluating it with the unseen test data after training [40]. The data can be split into three parts to refine this approach, introducing a validation set alongside the training and test data. The validation data makes evaluating the model during training possible by giving insight into how to further tune the model's hyperparameters and enhance the overall performance.

For the model developed and used for this thesis, the validation data is specified within the built-in method by Keras when instantiating the training process. Within this method, an argument enables the definition of the desired portion of training data allocated for use as a validation set. During the training phase, the model does not have access to this data segment but utilizes it for evaluation following each epoch.

After training, a graphic analysis of the training and validation loss is evaluated before tuning the model's hyperparameters to improve performance.

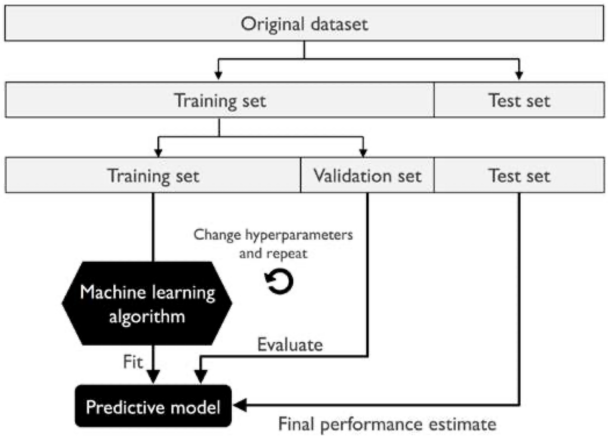
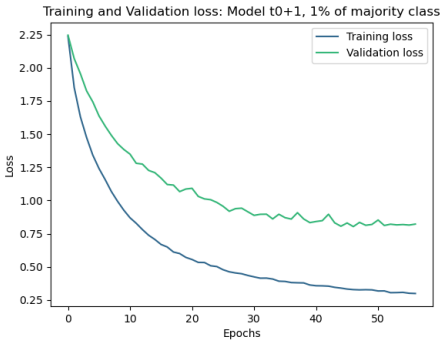



Figure 8: Illustration of how the concept of holdout cross-validation is performed. The original data is split into training and test data, where a portion of the training data is selected as a validation set used throughout the model training [38].

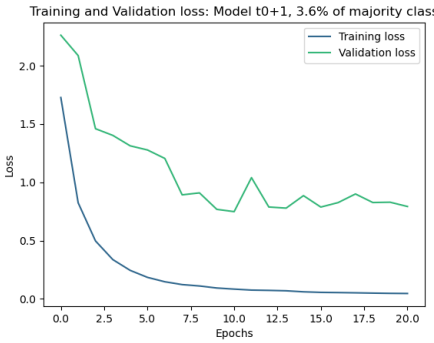
For the model developed and used for this thesis, the validation data is specified within the built-in method by Keras when instantiating the training process. Within this method, an argument enables the definition of the desired portion of training data allocated for use as a validation set. During the training phase, the model does not have access to this data segment but utilizes it for evaluation following each epoch. Figure 8 presents a visual representation of the concept of holdout cross-validation.

Post-training, an essential step in model refinement, involves thoroughly analyzing the training and validation loss. Visualizing the learning curves presented in Figure 9 enables conclusions about performance. Additionally, this method aids in identifying potential instances of underfitting and overfitting, which gives essential insights into how the model learns from the input data.

Throughout the model's training process, the F1 score has been implemented as an evaluation metric for both the training and validation datasets. In the subsequent section, more detailed information regarding the F1 score will be provided, which covers evaluation metrics for the test data.




(a) Example of learning curves for model $t_0 + 1$ with 1% of majority class.

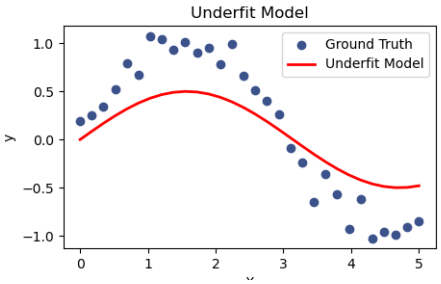



(b) Example of learning curves for model $t_0 + 1$ with 3.6% of majority class.

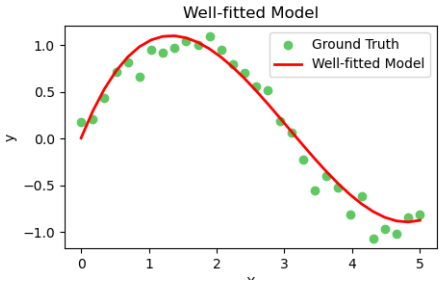
Figure 9: Plot of learning curves for the same model, but varied size of the majority class included in training. Same model and parameter values have been used for both.

#### 2.6.1.1 Model Fitting

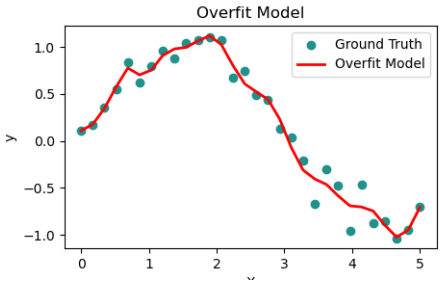
Acquiring the most optimal ML-model for the objective requires finding a balance among all the various aspects of the model that can be fine-tuned and explored. A well-fitted model demonstrates promising performance on the training data and its capacity to generalize effectively to new and unseen data, as evidenced by robust performance on the validation set, see Figure 10b. The balance between the performance on the training and validation set can reveal potential issues, and analyzing the learning curves from the training phase can show if the model is showcasing a good fit, or if it is underfitting or overfitting.




(a) Underfitted model: Not able to capture the complexity in the data.




(b) Well-fit model: Captures the most essential patterns in the data, creating a balanced fit.




(c) Overfitted model: Model captures noise and randomness in the data, will not regularize well on unseen data.

Figure 10: Comparison of how different model fits capture the data.

When a model suffers from underfitting like in Figure 10a, also referred to as having a high bias and low variance, the algorithm fails to adequately capture the underlying patterns in the data because of its lack of complexity. The bias is the difference between the predicted variable and the ground truth, and the variance describes how sensitive the model is to small changes in the input data. Consequently, the model is unable to learn the relationship between the independent and target variables, leading to a model that performs poorly on training and unseen data.

Underfitting, Figure 10c, can be detected visually when the training loss decreases over time while the validation loss remains consistently high or displays irregular spikes of relatively high values.

Increasing the complexity of the model is a typical solution to address underfitting. For ANNs, this can be achieved by experimenting with adding

more layers and increasing the number of units within each layer. Additionally, increasing the number of training epochs and, if available, adding more features to the input can provide the model with additional information. If regularization techniques are applied, these could be decreased in strength.

Conversely, overfitting implies that the model learns the structure of the input data too well. Overfitting means that potential noise or random instabilities are considered meaningful patterns by the model, decreasing the model's ability to generalize to new unseen data where this exact pattern does not exist. From a visual perspective, this scenario would present itself with training loss that decreases over time, while validation loss decreases to a point where it starts to increase again.

Since overfitting describes models that are low in bias and high in variance, regularization is a technique that increases the bias and decreases the variance to obtain better generalization from the model. There are several implementation strategies [46], and the ones used in this thesis are Batch Normalization and L1 regression.

### 2.6.2 Test Data

Following the training phase, the test dataset is introduced as input for the model to assess how it performs with new and unseen data. In the context of this thesis, various methods are employed for evaluating the test data, each offering additional insights into the model's overall performance and further enhancement.

#### 2.6.2.1 Confusion Matrix

A confusion matrix is a typical method for describing how well a model distinguishes between different class labels when dealing with a classification problem [38]. The information is presented through the four components in Figure 11 and represented in a tabular format, as shown in Table 2.

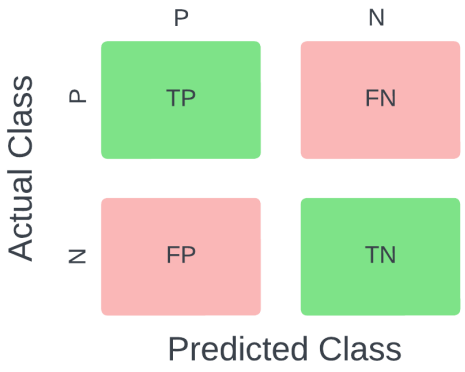



Figure 11: Representation of a confusion matrix.

Table 2: Description of the compontents making up the confusion matrix in Figure 11.

| **Component** | **Description** |
|---|---|
| *True Positives (TP)* | Predicted class is positive when actual label is positive |
| *True Negatives (TN)* | Predicted class is negative when actual label is negative |
| *False Positives (FP)* | Predicted class is positive when actual label is negative |
| *False Negatives (FN)* | Predicted class is negative when actual label is positive |

Various performance metrics, including accuracy (ACC), recall (REC), precision (PRE), and F1 score, can be calculated from the components of a confusion matrix. Accuracy is computed by taking the sum of all correctly predicted labels and dividing it by the sum of all predicted labels, presented in Equation 4 [38]. One can compute the same metric through the error rate (ERR) in Equation 3.

Given the imbalance in the data used in this thesis, it is essential to note that accuracy can be misleading for evaluation [5]. If the model performs well in classifying the majority class but poorly predicts the minority class,

it could still achieve a high accuracy score. Accuracy as a metric will then fail to provide a nuanced assessment of the overall performance. To address this limitation, evaluating the False Positive Rate (FPR) and (TPR) can better indicate the performance [38].

$$ERR = \frac{FP + FN}{FP + FN + T + TN} \tag{3}$$

$$ACC = \frac{TP + TN}{FP + TN + FN + FP} = 1 - ERR \tag{4}$$

$$PRE = \frac{TP}{TP + FP} \tag{5}$$

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP} \tag{6}$$

$$TPR = REC = \frac{TP}{P} = \frac{TP}{FN + TP} \tag{7}$$

In this thesis, the FPR would indicate the proportion of instances where non-icing weather phenomena are incorrectly classified as occurrences of icing weather. A low FPR score suggests that the model seldom misclassifies negative instances as positive, and overall, there are fewer false positive errors.

On the other hand, TPR, also called recall, measures the model's ability to correctly identify actual instances of icing weather phenomena from the entire set of recorded icing weather instances. The desired outcome of the ANN is to identify freezing weather phenomena at airports, which could be indicated by achieving a high TPR score that reflects the model's ability to identify the positive cases correctly. To this end, we also consider the F1 score:

$$F1 = 2\frac{PRE * REC}{PRE + REC}. \tag{8}$$

The F1 score in Equation (8) combines the strengths of precision and recall, offering a balanced measure by assessing the model's performance on false positives and false negatives [38]. With values ranging from one to zero, a higher F1 core signifies better performance, reflecting a balance between precision and recall. As mentioned in Section 2.6.1, this metric serves as a way to evaluate the model during training and further applied during the prediction phase for comparison.

#### 2.6.2.2 Receiver Operating Characteristic Curve

A Receiver Operating Characteristic (ROC) curve is a standard tool for model evaluation [38, 21]. It plots the TPR against the FPR at various thresholds for the specific classifier. The model is evaluated based on the ROC curve by computing the Area Under the ROC Curve (AUC), resulting in a value between zero and one. A value around 0.5 indicates that the model is just making random guesses, while an excellent model would achieve a value above 0.9. A fair to good model falls within the range of 0.7 to 0.9 [29].

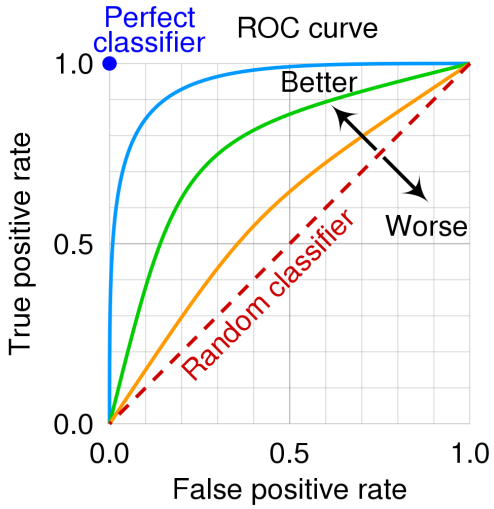



Figure 12: Visual representation of how a ROC curve could be plotted and how they differ in performance evaluation [39].

#### 2.6.2.3 Reliability Diagrams and Brier Score

A reliability diagram, often called a calibration curve, is another graphical tool for evaluating probabilistic classification models. In the context of this thesis, the ANN functions as a probabilistic classifier, predicting values between zero and one to represent the likelihood of icing weather occurrences at specific times. Assessing the model's calibration quality depends on how accurately these predicted probabilities align with the observed frequency of events [14]. The Brier score, closely related to the reliability diagram, serves as a measure of the calibration performance of probabilistic forecasts, commonly used in meteorology [17, 22]. It estimates the accuracy of probability forecasts, where a perfect forecast would yield a score of zero, while complete

inaccuracy would result in one. An ideal score approaches zero, indicating optimal performance.

# 3 Methodology

The Methodology section of this thesis provides a structured overview of the research approach, surrounding data collection, model selection, implementation, and prediction methodologies. First, the Data subsection gives an overview of the collection process, data quality assessment, and further feature selection. Subsequent sections cover Model Selection and Model Implementation and Development, also addressing issues such as over- and underfitting, regularization of the data, and model evaluation. The whole section is rounded off by detailing the prediction stage, exploring predictive techniques, examining instantaneous and future predictions, and finalizing them by using ROC as an evaluation metric for the result.

## 3.1 Data

### 3.1.1 Collection of Data

The data utilized in this study was collected from two different sources, both of which are part of MET's databases. The target variable we want to predict is the occurrence of the FZ in a METAR report, indicating the presence of freezing weather. As for the independent variables, the second part of the data consists of meteorological parameters with corresponding values aimed at predicting the target variable.

#### 3.1.1.1 Airports

This study employs data linked to the geographical locations of land-based airports across Norway. The location of each specific airport, defined by longitude and latitude coordinates, is derived from a PostgreSQL database at MET, which contains historical METAR reports from all airports in Norway, including offshore installations such as oil platforms and airports on Svalbard. The number of airports in the final dataset, which is further discussed in Section 3.1.4, is 54.

#### 3.1.1.2 METAR Reports

METAR reports serve as the ground truth in this study, also called the dependent variable, as the variable we want to predict. These reports document meteorological conditions observed at the airports included in the research. The content of the collected historical METAR reports gives detailed information about past conditions. As mentioned in Section 2.1.1, they are issued every 30 minutes, and in this case, at 20 minutes past the hour and 50 minutes to the hour.

The extraction process involved accessing a dedicated PostgreSQL database at MET, specifically designed to store METAR reports. Data from January 1, 2021, until December 31, 2023, was easily extracted from the database. To ensure the inclusion of exclusively Norwegian airports, identifiers beginning with 'EN' representing Europe Norway were specified. The extracted data contains the METAR report, the time of issuance, the airport identifier, and the airport location defined by longitude and latitude coordinates. Before preprocessing the data, the total collection of archived METAR reports was 3,405,525. Notably, in terms of message issuance, Kristiansund (ENKB), Kristiansand (ENCN), Sandefjord (ENTO), Hammerfest (ENHF), and Gardermoen (ENGM) emerged as some of the most prominent airports, with each contributing approximately 50,000 METARs each.

Analyzing the retrieved messages involves identifying instances containing the abbreviation "FZ," indicative of freezing weather incidents. Incorporating this information into the query reveals that Namsos (ENNM), Gardermoen (ENGM), Røros (ENRO), and Andøya (ENAN) are airports where freezing weather is most frequently forecasted. Comparing the count of METAR messages with FZ to the total METAR count yields the following statistics presented in Table 3.

Table 3: Information about registered METAR messages and FZ registration at specific airports from 2021-01-01 until 2023-12-31.

| **Airport Identifier** | **Total METAR** | **Total METAR with FZ** |
|---|---|---|
| Namsos (ENNM) | 35961 | 2324 |
| Gardermoen (ENGM) | 53059 | 1133 |
| Røros (ENRO) | 52357 | 623 |
| Sandefjord, Torp (ENTO) | 53056 | 569 |

The registered messages containing FZ account for 6 percent of the total

registered messages for Namsos. The remaining three messages with FZ account for 1-2 percent of the total.

As highlighted in Section 1.3.2, the issue of faulty data, mainly originating from airports with auto-generated METAR reports, is a potential concern. Auto-generated messages are known to contain inaccuracies and wrongly forecast freezing weather events. In complement to the data outlined in Table 3, an additional query is executed to assess the frequency of occurrences where auto-generated METAR messages include the abbreviation FZ.

Table 4: Information about auto-generated METAR messages and FZ registration at specific airports from 2021-01-01 until 2023-12-31.

| **Airport Identifier** | **Total AUTO METAR with FZ** |
|---|---|
| Namsos (ENNM) | 2225 |
| Gardermoen (ENGM) | 0 |
| Røros (ENRO) | 438 |
| Sandefjord, Torp (ENTO) | 205 |

At Gardermoen, there are no auto-generated METAR messages containing FZ. However, in two of the three remaining airports, the number of auto-generated METAR messages with FZ exceeds the manually forecasted count. For Namsos, it accounts for almost all the instances. Further observation shows that, when factoring in the geographical locations of the mentioned airports, there is no indication that freezing weather incidents are more prevalent in one location than another based solely on registered messages.

Since the METAR reports are issued every 30 minutes, precisely at 20 minutes past and to the hour, adjustments to timestamps were necessary to align them with the meteorological parameters' whole-hourly timestamps. The method for addressing this involved rounding off the timestamps for issuance to the whole hour, which resulted in several reports with identical timestamps. The first step in handling these duplicates was to create a priority list of the available METAR types, with SPECI, COR, MANUAL, and AUTO in their respective order. The data was then sorted based on this priority list and the timestamp for issuance. Duplicates were removed by retaining the first occurrence of each METAR type according to the sorted list.

#### 3.1.1.3 Meteorological Parameters and Data Processing

The meteorological parameters are selected in cooperation with and following discussions with meteorologists at MET. Collaborating with professionals who deal with this daily ensures that the decision is grounded in solid insights, further enhancing the foundation for the study and reliability of the attempted forecasts and predictions presented later in this thesis.

Determining freezing weather involves analyzing various factors. While a temperature reaching the freezing point is needed to cause the formation of ice or frost and stands as an essential contributor, the interaction of additional parameters is equally significant. These factors collectively cause diverse scenarios of freezing weather. Table 5 presents all the chosen parameters extracted from MET's databases, displaying the parameter name, unit, and brief description.

The final dataset includes all of the parameters presented, with some modifications. Parameters with Kelvin as the unit, such as *air_temperature_0m*, *air_temperature_2m*, *air_temperature_pl_850*, and *air_temperature_pl_925*, have been converted to Celsius. Furthermore, the accumulated precipitation amount, denoted by the parameter *precipitaion_amount_acc*, has been substituted with a parameter representing the hourly calculated precipitation amount.

Table 5: Extracted parameters from MEPS data.

| **Parameter** | **Unit** | **Description** |
|---|---|---|
| air_temperature_0m | Kelvin | Surface temperature (T0M) |
| air_temperature_2m | Kelvin | Screen level temperature (T2M) |
| air_temperature_pl_850 | Kelvin | Temperature 1.5 km above sea level |
| air_temperature_pl_925 | Kelvin | Temperature 750-800m above sea level |
| relative_humidity_2m | | Screen level relative humidity (RH2M) |
| precipitation_amount_acc | $kg/m^2$ | Accumulated total precipitation |
| x_wind_10m | m/s | Horizontal component of wind 10m above sea level |
| x_wind_pl_850 | m/s | 1.5km above sea level |
| x_wind_pl_925 | m/s | 750-800m above sea level |
| y_wind_10m | m/s | Meridional 10 metre wind (V10M) |
| y_wind_pl_850 | m/s | 1.5 km above sea level |
| y_wind_pl_925 | m/s | 750-800m above sea level |
| surface_air_pressure | Pa | Surface air pressure, height=0 |
| air_pressure_at_sea_level | Pa | Mean Sea Level Pressure |
| fog_area_fraction | % | Ratio of fog coverage |
| liquid_water_content_of_surface_snow | $kg/m^2$ | Snow Water Equivalent |

The data extraction process involved accessing the databases at MET. Data produced by the MEPS model is stored in a NetCDF (network Common Data Form) format. This format can store multidimensional scientific data and can contain multiple arrays of shape and size, making it optimal for meteorological variables.

Each NetCDF file contains various parameters, all of which share the standard dimensions time, x, and y, alongside additional descriptors like height and pressure. A custom Python function was employed to traverse the folder structure in the database and pinpoint the necessary files for extraction. This function enabled identifying and extracting the desired parameters from 2021, 2022, and 2023. This process employed the registered location for each airport, accessible within the same database as the METAR reports, in longitude and latitude measurements. The extracted data corresponds to the nearest coordinates with parameter values for each airport.

Due to the size of the extracted data and the time-consuming nature of the process, this process was carried out individually for each month within

each year. The extracted data, in turn, was saved in pickle files, a choice made for the same practical reasons. Following this, the data underwent further processing and conversion into Pandas format, subsequently saved as CSV files, and finally concatenated to consolidate the dataset. This approach facilitates storage in a two-dimensional format, optimizing accessibility and efficiency for subsequent analysis and manipulation.

Observation of the extracted data revealed that certain periods and timestamps exhibited missing values for specific parameters. To address this, the chosen approach was to use forward linear interpolation to address all instances of missing data across all parameters. This method estimates and fills the missing values by assuming a linear relationship between neighboring instances and utilizing the subsequent available data points in a forward direction [38, 4].

### 3.1.2 Data Cleansing and Examination

The data examination process was conducted with a specific focus on freezing weather occurrences and key meteorological parameters that could serve as reliable indicators of these events. This involved a deeper exploration of the relationship between temperature and observed freezing weather at each airport. Figure 13 illustrates the highest and lowest registered temperatures across all airports for registered METAR reports containing FZ. Specifically, the highest and lowest recorded temperatures within these specifications were $5.68°C$ and $-28.51°C$, respectively. Based on these findings, all temperature values above $6°C$ were excluded from the final dataset. This decision to remove unnecessary data points will streamline analysis and ensure the data is better prepared for the models.

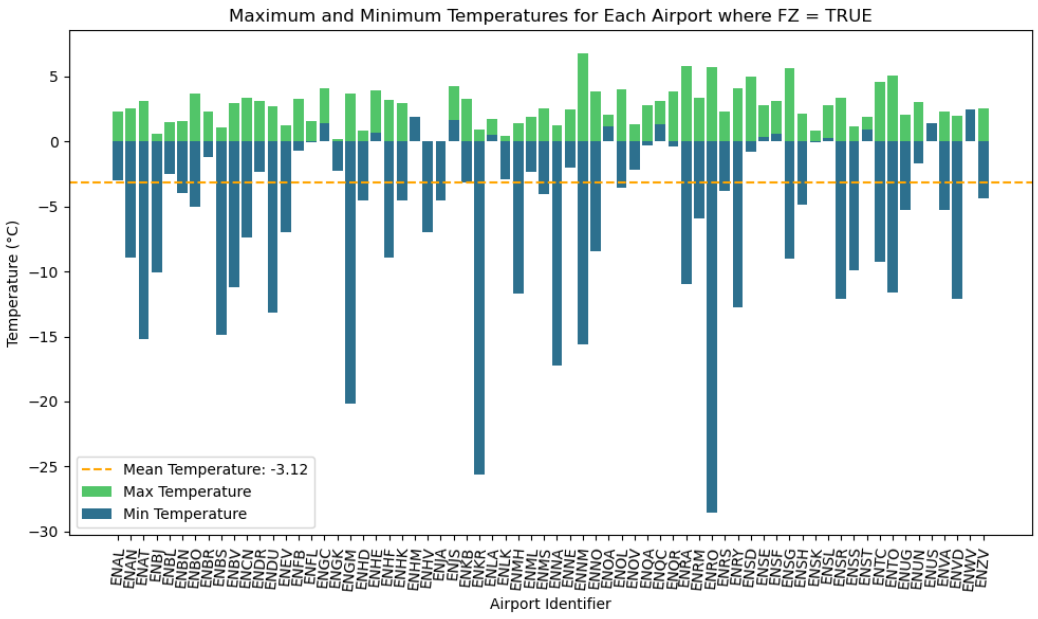



Figure 13: A bar diagram of maximum and minimum registered temperatures for observed freezing weather occurrences at each airport.

An analysis of the distribution of METAR reports containing freezing weather indicators throughout the year revealed a notable contrast among the months, as outlined in Table 6. Notably, due to higher temperatures, a decrease in instances is naturally observed during May, June, and August, with no reported occurrences in July. The information from these months was not included in the final dataset to retain only the most suitable data.

Table 6: Overview of the distribution of recorded FZ occurrences in METAR reports during the period 2021-01-01 until 2023-12-31.

| **Month** | **METAR containing FZ** |
|---|---|
| January | 1062 |
| February | 900 |
| March | 1034 |
| April | 403 |
| May | 84 |
| June | 12 |
| July | 0 |
| August | 10 |
| September | 49 |
| October | 307 |
| November | 758 |
| December | 1274 |

### 3.1.3 Data Manipulation

Following the process of cleaning and examining the data of the two extracted datasets, a few steps remain before merging them into a final dataset and initiating the implementation of the ANN.

The ground truth serves as the reference, which is compared to the predicted target variable from the model to further evaluate the model's performance. In this thesis, the information derived from the extracted METAR data serves as the ground truth, specifically regarding the occurrence of icing events at the locations of aerodromes.

The focal point of the ANN in this thesis is to predict the possibility of icing at aerodrome locations. This dependent variable relies on the presence or absence of the abbreviation "FZ" in reported METARs. A boolean variable is added in its separate column, flagging METAR messages that report either freezing rain (FZRA), freezing drizzle (FZDZ), or freezing fog (FZFG). These flags also include FZUP, indicating an undefined precipitation.

When predicting several timesteps ahead of the current one, the model requires input data comprising information from the current timestep up to the one just before the prediction horizon. For instance, if we aim to forecast 6 hours into the future, the input dataset must encompass observations from the current timestep ($t_0$) and subsequent timesteps up to $t_0 + 5$. To prepare

the dataset for this type of prediction, each parameter was extended with extra columns representing values for $n$ timesteps ahead of the current observation and the values for each parameter. Using *air_temperature_0m* as an example, the original parameter value at $t_0$ serves as the observation at this moment, with additional columns appended to account for $t_0+1, t_0+2, ..., t_0+5$.

The meteorological parameters are scaled through standardization before feeding this data to the models, ensuring uniform properties and distribution with a mean of zero and a standard deviation of one [38].

### 3.1.4 Final Dataset

After preprocessing, examining, and manipulating the data, the final dataset comprises 555313 rows $\times$ 126 columns. It includes meteorological parameters and METAR reports for the location of 54 unique airports from January 1, 2021, until December 31, 2023. Among these instances, the total number of freezing weather observed and documented in METAR reports, denoted by the FZ abbreviation, is 5,313.

## 3.2 Model Implementation and Development

Implementing and developing the ANN model for this project is completed through continuous testing and refinement of hyperparameters to achieve optimal performance. This iterative optimization process encompasses a multifaced exploration, including adjusting various architectural components and applying essential regularization techniques to balance complexity and performance.

Exploring different configurations of hidden layers and the number of units within each layer and systematically evaluating the learning curves from the training and validation data to indicate the model's status has been the most significant source of consistently improving the model little by little. Downsampling the majority class in the training data resulted in a relatively small dataset for training and validation. The reduced dataset enabled swift training iterations, allowing for efficient exploration of various model configurations and experimentation with different hyperparameters. However, the downside of having a restricted amount of training data is that it limits the availability of sufficient information for the model to learn from. This will compromise the model's ability to capture the full complexity of the underlying patterns in the data.

The following section will present the architecture of the final network, representing the completion of efforts within the available time constraints. This composition reflects the structure and configurations that yielded the most promising results during the training and testing.

### 3.2.1 Model Complexity and Architecture

Refining the ANN persisted until additional hyperparameter tuning ceased to produce noticeable improvements, signifying the conclusion of further architectural refinement. Figure 14 presents the final architecture of the ANN model.

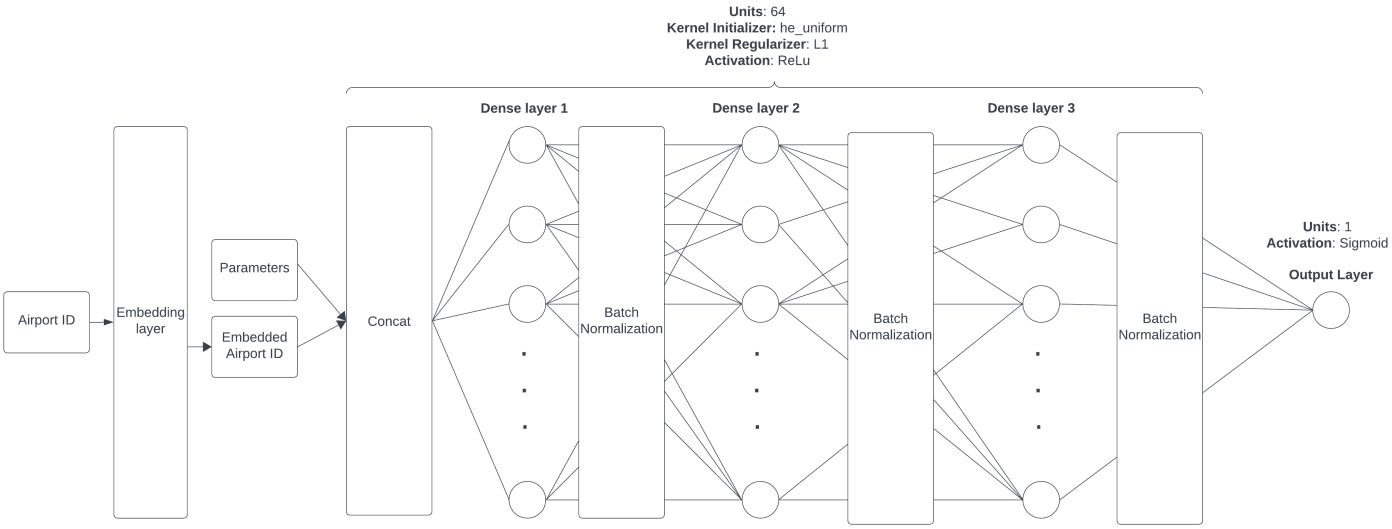



Figure 14: An overview of the final architecture of the ANN.

Each unique categorical airport identifier is assigned a numerical airport ID. Within the model's input processing stage, an embedding layer first converts the airport ID into a predetermined fixed-length numerical representation in vector form [38]. This transformation enhances the ML model's efficiency in handling categorical data. The embedded airport IDs are concatenated with the other inputs: the meteorological parameters and corresponding observations of icing weather occurrence.

The model contains two hidden layers alongside one input and output layer. Each layer depicted in Figure 14 is dense, indicating that every neuron in one layer connects to every neuron in the subsequent layer. Fewer hidden layers proved insufficient for effective information processing, while additional layers introduced needless complexity, even with heightened regularization strength. 64 units allocated to each layer achieved optimal performance. This was the sweet spot for balancing the model between capturing enough of the data patterns and avoiding excessive complexity, which could lead to overfitting. Alterations to the number of units caused similar outcomes to adjustments in hidden layers.

In addition to maintaining an identical number of units, each layer, excluding the output layer, sticks to the exact specifications concerning the initializer, regularizer, and activation function. In light of the binary classification task, the output layer is structured with a single unit and employs the Sigmoid activation function. Section 2.5.2 elaborates on the motivation behind these choices, providing insights into the selection process for activation functions.

The model architecture applies batch normalization between each layer. Initially, the testing rounds only featured dropout between the layers. However, performance improved upon experimenting with a combination of dropout and batch normalization. In the end, the best results were achieved solely by implementing batch normalization.

When compiling the model, a learning rate of 0.001 is assigned to the Adam optimization algorithm, employing binary crossentropy as the loss function and the F1 score as a metric. Increasing or decreasing the learning rate did not improve the model's performance. Additionally, early stopping is initialized, configuring it to monitor the validation set with patience of ten epochs. This setup ensures that training will halt if there is no improvement in performance for ten consecutive epochs.

The batch size is set to 32, and the number of epochs is 100. With early stopping implemented, training will cease when no further improvement is noticeable.

## 3.3 Probabilistic Classifier

Once the model has been defined, compiled, and fitted, it transitions into the training phase. Figure 15 provides an overview of the data structure encircling the meteorological parameters and how it is employed as input

Table 7: Overview of the hyperparameters and corresponding values.

| **Hyperparameter** | **Value** |
|---|---|
| L1 Regularization | 0.002 |
| Adam | 0.001 |
| Batch Size | 32 |
| Epochs | 100 |

for the model depending on the predictive task. These features are then concatenated with the embedded airport ID and corresponding observation of icing weather occurrences before it is ready for the model. See Section 3.2.1 and Figure 14 for more information.

The dataset is divided into training and test sets with an 80/20 ratio. When fitting the model, a validation split of 20% from the training set is specified to evaluate performance during this phase. The F1 score is also implemented as a metric alongside the training and validation loss.

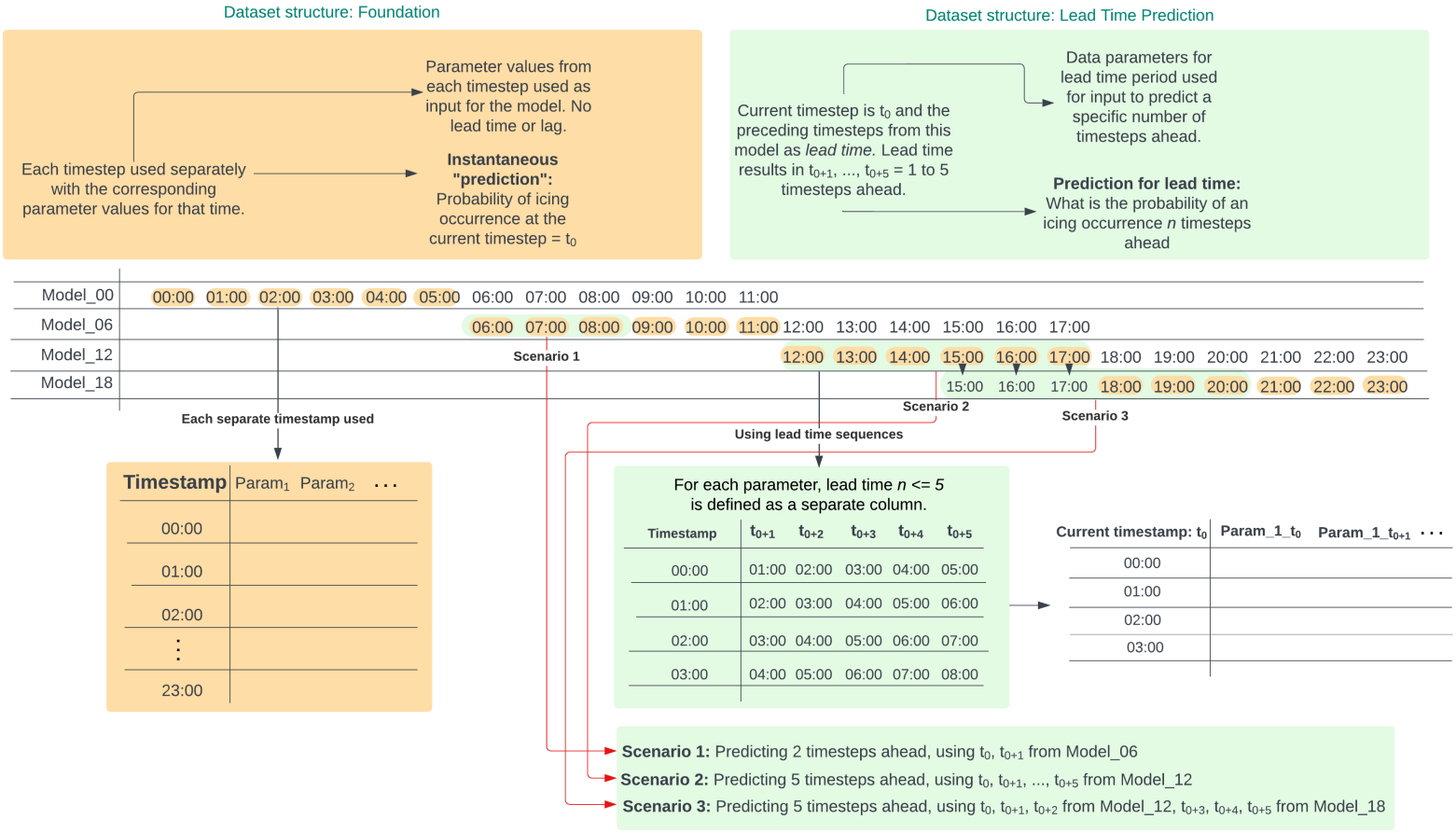



Figure 15: Detailed overview of how the extracted data is utilized in the prediction process, both as instantaneous prediction and predicting $n$ timesteps ahead.

Table 8: Overview of the different datasets with identifiers for easier referencing.

| **Dataset** | **Identifier** |
|---|---|
| 1% of majority class | D1 |
| 3.6% of majority class | D2 |
| All AUTO metartypes removed and classes balanced | D3 |

Table 9: Overview of the different predictive classifiers and datasets used together with the specific input.

| **Model** | **Prediction** | **Dataset** | **Details** | **Model Input** |
|---|---|---|---|---|
| 1 | Instantaneous | D1, D2, D3 | Current timestep | Embedded airport ID, and meteorological parameters. |
| 2 | 1H ahead | D1, D2, D3 | One timestep | Embedded airport ID, meteorological parameters and FZ observations for $t_0, t_{(0+1)}$ |
| 3 | 2H ahead | D1, D2 | Two timesteps | Embedded airport ID, meteorological parameters and FZ observations for $t_0, t_{(0+1)}, t_{(0+2)}$ |
| 4 | 3H ahead | D1, D2 | Three timesteps | Embedded airport ID, meteorological parameters and FZ observations for $t_0, t_{(0+1)}, t_{(0+2)}, t_{(0+3)}$ |
| 5 | 4H ahead | D1, D2 | Four timesteps | Embedded airport ID, meteorological parameters and FZ observations for $t_0, t_{(0+1)}, ..., t_{(0+4)}$ |
| 6 | 5H ahead | D1, D2 | Five timesteps | Embedded airport ID, meteorological parameters and FZ observations for $t_0, t_{(0+1)}, ..., t_{(0+5)}$ |
| 7 | Persistence | D1, D2, D3 | Current timestep | Embedded airtport ID |

After identifying the architecture and parameter values that produced the most promising results, consistency was maintained by keeping the foundational parameters for each model while exploring various datasets and inputs. Table 5 explains the different datasets, while Table 7 provides an overview of the different types of prediction, what datasets have been used, and the specific inputs utilized.

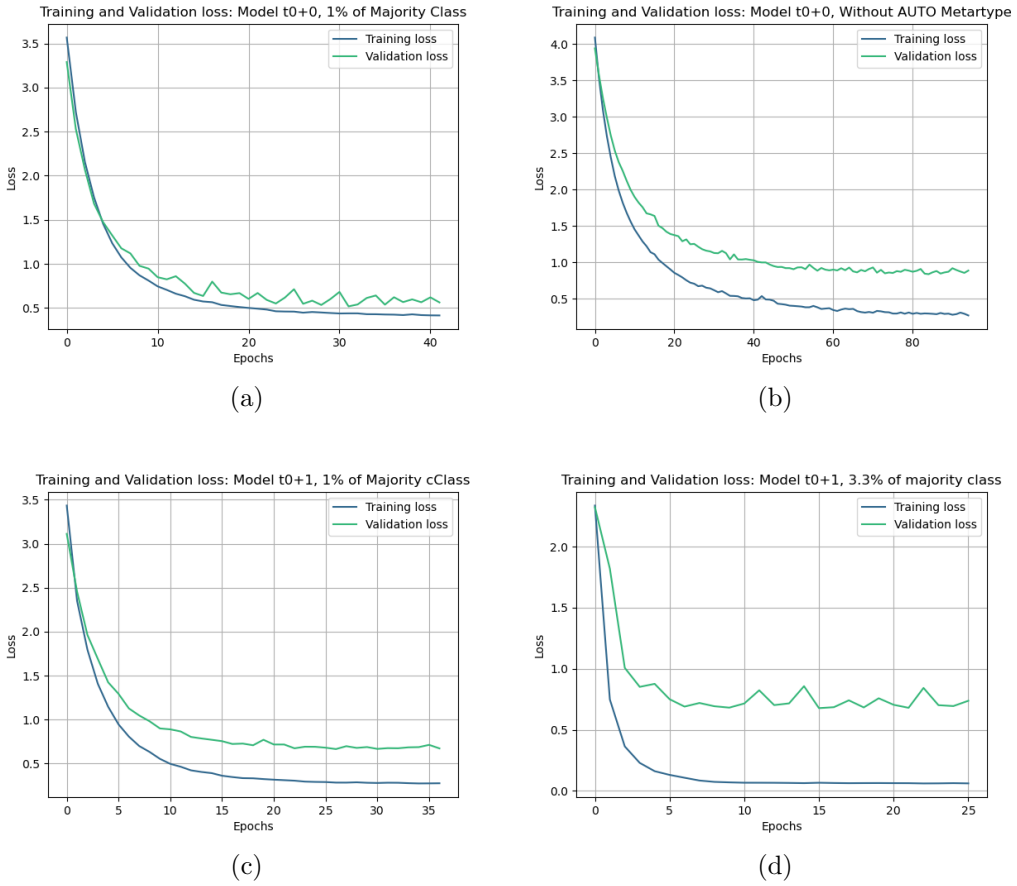



Figure 16: Training and validation plots showcasing four different data inputs.

Figure 16 illustrates four training and validation plots with three different datasets and four distinct types of inputs. In plots Figure 16a and Figure 16c, the same dataset, containing 1% of the majority class, is utilized for training. The plot in Figure 16a involves instantaneous prediction, whereas

the plot in Figure 16c predicts one timestep. In Figure 16b, the dataset is modified by removing the AUTO metartype, further balancing the majority and minority class with a difference of 300 instances more of the majority. Finally, Figure 16d illustrates the prediction of one timestep using a dataset containing 3.3% of the majority class.

The remaining models have not been included as their plots closely resemble Figure 16c for the dataset with 1% of the majority class and Figure 16d for the dataset with 3.3% of the majority class.

The number of epochs displayed on the x-axis varies depending on when early stopping has interfered with and stopped the training. An ideal plot showcasing a well-fit model would exhibit both training and validation loss decreasing concurrently and plateauing where the validation can be higher than the training loss. This behavior is observable in Figure 16a.

In Figures 16b and 16c, a similar pattern emerges, with both training and validation loss decreasing before leveling off. However, the model performs notably better on the training data than the validation data, suggesting a reduced ability to generalize to new and unseen data. This observation suggests that the model is overfitting slightly.

These results arise from the final version of the model architecture and hyperparameter values that offer the best overall performance after the iterative phase of training and hyperparameter tuning. Figure 14 and Table 7 contain the details about the final architecture and hyperparameter values used.

## 3.4 Code

The code used for the work and experiments in this thesis is available for reference, ensuring reproducibility of the models presented. Additionally, high-resolution images of the figures and plots included throughout this thesis are provided for clarity and detailed examination.
`https://github.com/tomaloki/2024_msc_tonje_metar`

## 3.5 Usage of Additional Tools in the Thesis Writing

In crafting this thesis, the author has employed various tools to enhance the flow and diversity of the language. Given that English is the author's second language, additional support and guidance were sought through specialized

AI-based software, notably Grammarly[1] and ChatGPT[2]. These tools have helped identify grammatical errors, enhance formulation, and suggest alternative phrasings, contributing to the overall clarity and effectiveness of the writing.

The majority of figures presented in this thesis were created by the author using a software program called Lucidchart[3]. This tool facilitated the visualization of complex concepts, data, and architecture, enabling better explanation and representation throughout the thesis.

# 4 Results

This section presents the outcomes of the model evaluation following the completion of the final test phase and analyzes the final results and predictions. The test dataset, comprising 20% of the total data obtained through the test and training split, remains unaltered, unlike the training data, and maintains its inherent imbalances.

An overview of the dataset reveals 111,063 instances, with a mere 1070 freezing weather events observed. The small size of the minority class highlights the significant skewness in the data, posing a notable challenge in accurately assessing the model's efficiency in predicting freezing weather conditions, especially given the small size of the better-balanced training set.

It is important to note that dataset D3, see Table 6 for reference, has been excluded from training in the temporal models. However, it has been retained for comparison in the persistence and instantaneous models. The removal of all AUTO METAR reports resulted in an even smaller portion of the minority class compared to the majority. Considering the size of the data and the time constraints, the decision was made to exclude dataset D3 from training the temporal models.

## 4.1 Prediction

This study categorizes the predictions into two subsections. The first method, instantaneous, predicts the likelihood of freezing weather occurring at the current timestep. The second method involves predicting freezing weather

[1] `https://www.grammarly.com/`
[2] `https://www.chatgpt.com/`
[3] `https://www.lucidchart.com/`

occurrences $n$ timesteps ahead, aiming to provide accurate predictions of freezing weather possibility in the near future.

A persistence model serves as an additional baseline and is trained and evaluated differently from conventional models. Instead of training to learn patterns and make predictions, the persistence model remembers the last observed target value and assumes that the current condition will persist into the future [35, 32]. For instance, if we observe icing weather today, the persistence model assumes the exact condition will occur in the following defined timestep.

The persistence models have undergone testing using the datasets detailed in Table 8. All three models, presented in Figure 17, demonstrate relatively high values upon examining the ROC curve and corresponding AUC scores. However, the AUC-PR scores reside on the lower end of the scale. A desirable plot would feature a larger area under the curve for the precision-recall plot, resulting from high precision and recall.

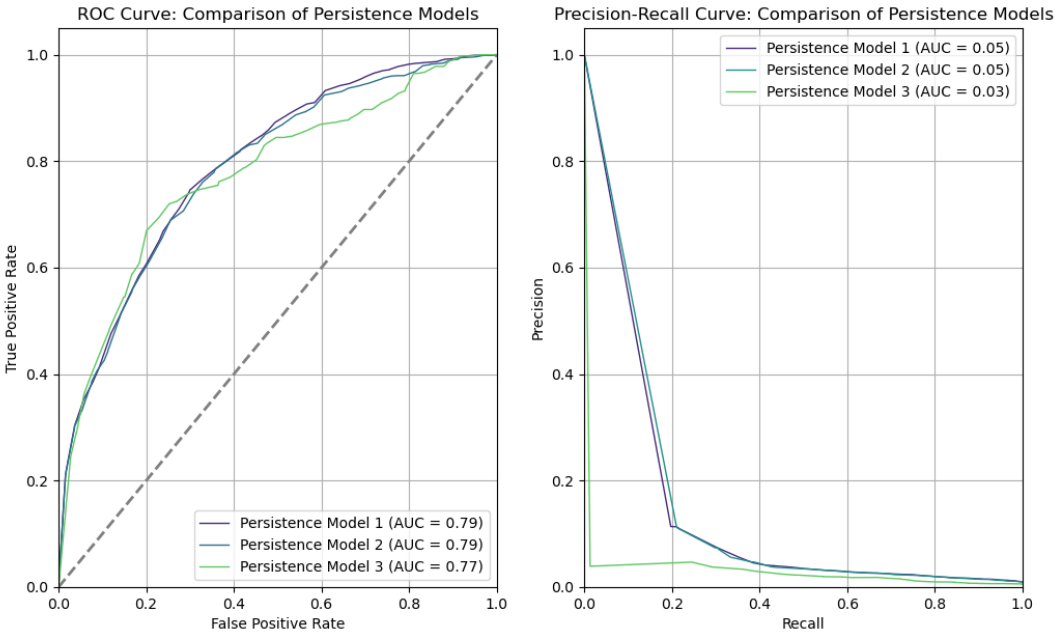



Figure 17: Persistence model for dataset D1, D2, and D3 and its resulting ROC Curve and Precision-Recall Curve with corresponding AUC and AUC-PR scores.

Precision tends to be higher for all three models at lower recall values. However, there is a notable drop as recall increases, particularly evident beyond a recall of 0.2 for models 1 and 2. Model 3 experiences an immediate decline in precision. These observations are further detailed in Table 10 which presents the precision, recall, and F1 scores across the thresholds 0.1, 0.4, and 0.8, which will be used throughout the evaluation of all models. Models 2 and 3 had notably no probability values above the 0.8 threshold.

Table 10: Overview of precision, recall, and F1 values for the three persistence models at three different thresholds.

| Model | Threshold | Precision | Recall | F1 |
|---|---|---|---|---|
| Persistence Model 1 | 0.1 | 0.01 | 0.99 | 0.02 |
| Persistence Model 1 | 0.4 | 0.02 | 0.74 | 0.05 |
| Persistence Model 1 | 0.8 | 0.03 | 0.54 | 0.06 |
| Persistence Model 2 | 0.1 | 0.03 | 0.62 | 0.05 |
| Persistence Model 2 | 0.4 | 0.11 | 0.21 | 0.15 |
| Persistence Model 2 | 0.8 | - | - | - |
| Persistence Model 3 | 0.1 | 0.01 | 0.83 | 0.02 |
| Persistence Model 3 | 0.4 | 0.03 | 0.37 | 0.06 |
| Persistence Model 3 | 0.8 | - | - | - |

The low AUC score in the precision-recall plot underscores that these models were trained solely on observed data without additional parameters as input features, implying there is no additional data for the models to learn from. These results will serve as a baseline for comparison, with the expectation that the remaining trained models will ideally surpass their performance.

### 4.1.1 Instantenous Model

The instantaneous model predicts the probability associated with the current temporal state. Unlike prognostic models that conclude into the future, it limits its analysis solely to the immediate present. As such, the outputs of an instantaneous model could serve as a verification or indication of the forecast for the specific hour and offer insights into the immediate state of affairs.

Table 9 presents the datasets used alongside the selected input for the model. Notably, these three models are only provided with values for the meteorological parameters at the specific timestamps and the embedded airport ID as input and independent variables. In contrast, the observational METAR data, serving as the target variable, is solely used as the variable we want to predict.

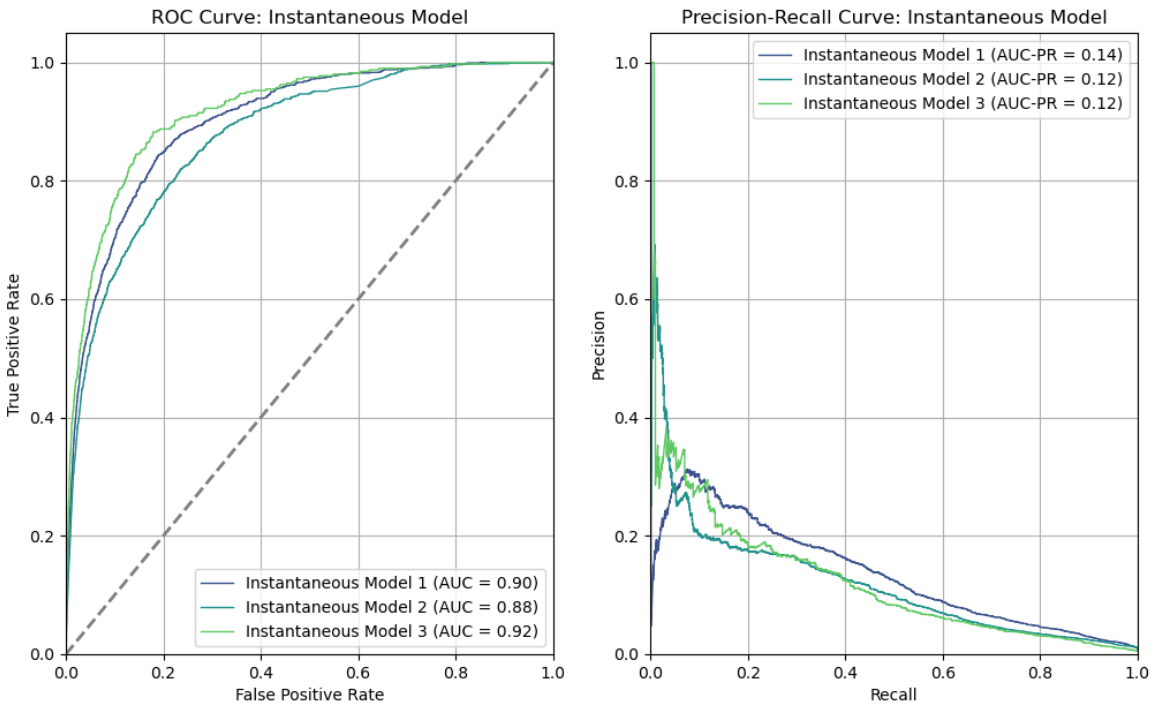



Figure 18: Instantaneous model for dataset D1, D2, and D3 and its resulting ROC Curve and Precision-Recall Curve with corresponding AUC and AUC-PR scores.

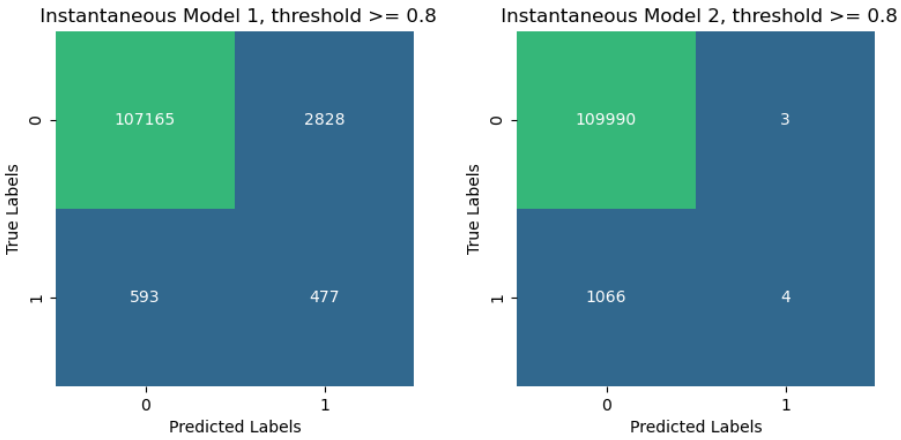
All three models, plots depicted in Figure 18, deliver high values for the ROC AUC, with two at or exceeding 0.90. However, the AUC-PR values are notably lower, ranging between 0.12 and 0.14, indicating a suboptimal precision-recall trade-off. The models struggle to maintain precision at higher recall thresholds, where the challenge is accurately predicting true positives. Consequently, while these models successfully capture positive instances, they do so at the expense of increased false positives, where negative cases are incorrectly classified as positive.

Table 11: Overview of precision, recall, and F1 values for the three instantaneous models at three different thresholds.

| **Model** | **Threshold** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|
| Instantaneous Model 1 | 0.1 | 0.02 | 0.97 | 0.04 |
| Instantaneous Model 1 | 0.4 | 0.04 | 0.84 | 0.08 |
| Instantaneous Model 1 | 0.8 | 0.14 | 0.45 | 0.22 |
| Instantaneous Model 2 | 0.1 | 0.04 | 0.73 | 0.08 |
| Instantaneous Model 2 | 0.4 | 0.16 | 0.30 | 0.21 |
| Instantaneous Model 2 | 0.8 | 0.57 | 0 | 0.01 |
| Instantaneous Model 3 | 0.1 | 0.01 | 0.84 | 0.02 |
| Instantaneous Model 3 | 0.4 | 0.03 | 0.40 | 0.06 |
| Instantaneous Model 3 | 0.8 | - | - | - |

All three models, plots depicted in Figure 18, deliver high values for the ROC AUC, with two at or exceeding 0.90. However, the AUC-PR values are notably lower, ranging between 0.12 and 0.14, indicating a suboptimal precision-recall trade-off. The models struggle to maintain precision at higher recall thresholds, where the challenge is accurately predicting true positives. Consequently, while these models successfully capture positive instances, they do so at the expense of increased false positives, where negative cases are incorrectly classified as positive.

This observation is reflected in the precision, recall, and F1 scores for the three selected thresholds in Table 9. This trade-off highlights that the model struggles to accurately identify positive instances while minimizing false positives.




(a) (b)

Figure 19: Confusion matrix showing the predictions from the Instantaneous models 1 and 2 with a threshold $>= 0.8$.

One interesting observation in Table 9 pertains to the values for Model 2 at a threshold of 0.8. The resulting recall value is 0, but the precision is 0.57. To explain these values, we need to observe the resulting confusion matrix in Figure 19b.

Using the formulas for precision and recall from Equations (5) and (7) in Section 2.6.2.1, we have the following:

$$PRE = \frac{TP}{TP + FP} = \frac{4}{4 + 3} = 0.57 \tag{9}$$

$$REC = \frac{TP}{P} = \frac{TP}{FN + TP} = \frac{4}{1066 + 4} = 0.0037 \tag{10}$$

Because of the high imbalance between false negatives and true positives, the recall value is reduced to near zero and is rounded off.

Figure 19a shows a confusion matrix of Model 1 at the same threshold for comparison. This model achieves recall and precision values of 0.14 and 0.45, respectively. Compared to Model 2, Model 1 has a significantly lower number of false negatives but a higher count of true positives. Model 1 also has a higher number of false positives.

### 4.1.2 Temporal Model

The temporal models forecast the probability of freezing weather occurrences from 1 to 5 timesteps ahead, where one timestep equals one hour. Individual models are trained for each specified timestep, and the input features are adjusted accordingly to accommodate the desired prediction horizon. Table 9 in Section 3.3, presents a detailed overview of the input configurations for each temporal model.

The results obtained from the models trained on dataset D1 exhibit diverse performance levels across various timesteps and thresholds. Analysis of the plots alongside the precision, recall, and F1 values in Table 12 illustrates a spectrum of effectiveness in predicting positive instances while concurrently minimizing false positives.

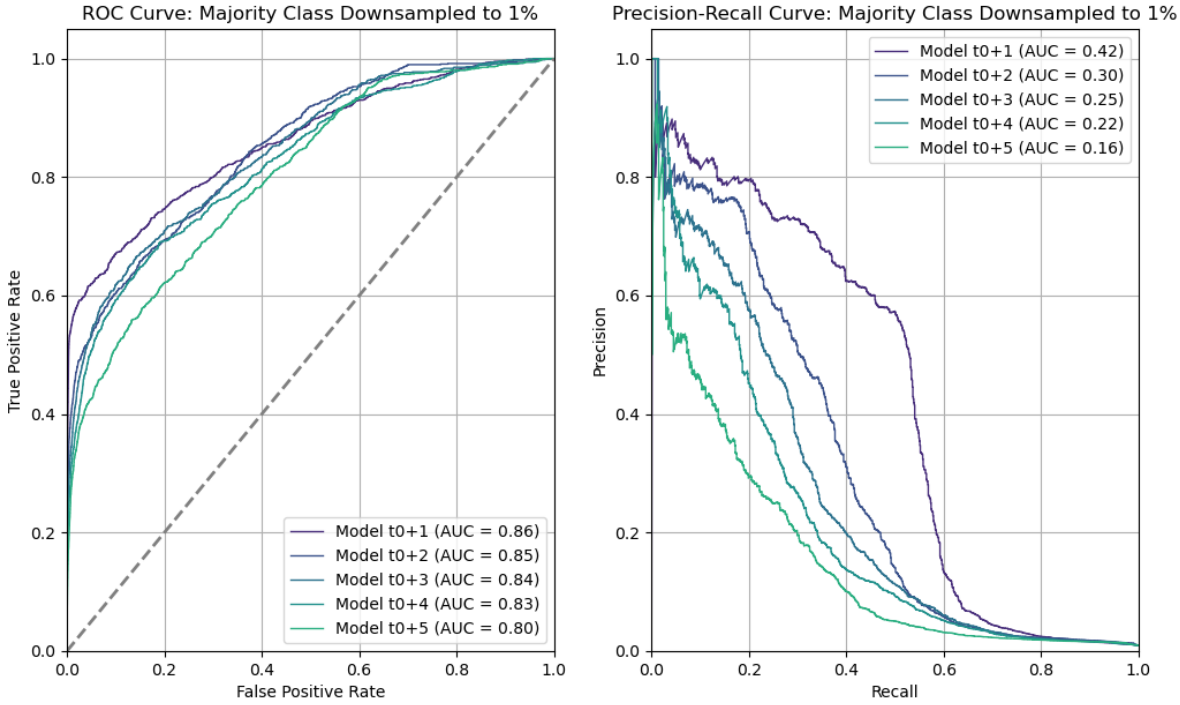



Figure 20: ROC and Precision-Recall plot of the temporal model trained on dataset D1. In Table 6 this includes Model 2-6.

To better illustrate the witnessed performance, we can compare the results in Table 12 with Tables 10 and 11. With the temporal models, there are consistently better precision values and an increased balance between precision and recall values. The overall F1 scores are higher, underlining a better balance between these metrics.

Table 12: Overview of precision, recall, and F1 values for the instantaneous model trained with dataset D1.

| **Model** | **Dataset** | **Threshold** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|---|
| $t_0 + 1$ | D1 | 0.1 | 0.37 | 0.55 | 0.44 |
| $t_0 + 2$ | D1 | 0.1 | 0.27 | 0.47 | 0.33 |
| $t_0 + 3$ | D1 | 0.1 | 0.25 | 0.35 | 0.29 |
| $t_0 + 4$ | D1 | 0.1 | 0.23 | 0.32 | 0.27 |
| $t_0 + 5$ | D1 | 0.1 | 0.13 | 0.36 | 0.19 |
| $t_0 + 1$ | D1 | 0.4 | 0.62 | 0.41 | 0.50 |
| $t_0 + 2$ | D1 | 0.4 | 0.64 | 0.22 | 0.33 |
| $t_0 + 3$ | D1 | 0.4 | 0.60 | 0.19 | 0.29 |
| $t_0 + 4$ | D1 | 0.4 | 0.58 | 0.15 | 0.23 |
| $t_0 + 5$ | D1 | 0.4 | 0.38 | 0.15 | 0.21 |
| $t_0 + 1$ | D1 | 0.8 | 0.83 | 0.09 | 0.16 |
| $t_0 + 2$ | D1 | 0.8 | 0.90 | 0.02 | 0.03 |
| $t_0 + 3$ | D1 | 0.8 | - | - | - |
| $t_0 + 4$ | D1 | 0.8 | 1.00 | 0.01 | 0.02 |
| $t_0 + 5$ | D1 | 0.8 | - | - | - |

As the prediction horizon extends across the five models forecasting for different timesteps, model performance varies noticeably. The trend here is diminishing precision and recall values, where the model struggles more to predict positive instances as the forecast horizon lengthens.

In the ROC plot, an optimal model typically exhibits a curve closer to the top-left corner, suggesting outstanding discriminatory ability. Achieving a higher TPR before the curve deviates from the y-axis signifies that the model is better at discriminating between positive and negative instances. As we extend the prediction horizon, this deviation occurs earlier, underscoring the model's ability to anticipate outcomes over longer timeframes.

Although all five models yield AUC scores surpassing 0.80, with the highest at 0.87, caution is necessary due to the class imbalance in the test data.

High AUC scores may obscure performance issues, particularly concerning the minority class, where the model may struggle despite accurately classifying the majority. The precision-recall curve is an important addition here, providing more insights into the performance of the models across different thresholds, particularly in scenarios with imbalanced data.

The model for $t_0+1$, achieving an AUC of 0.86, demonstrates a robust discriminatory power in distinguishing between positive and negative instances. However, with an AUC-PR of 0.42, it only exhibits moderate performance in capturing true positives without introducing too many false positives.

Considering the four remaining models showcasing an AUC score exceeding 0.80, a notable decline in AUC-PR is evident for all remaining timesteps.

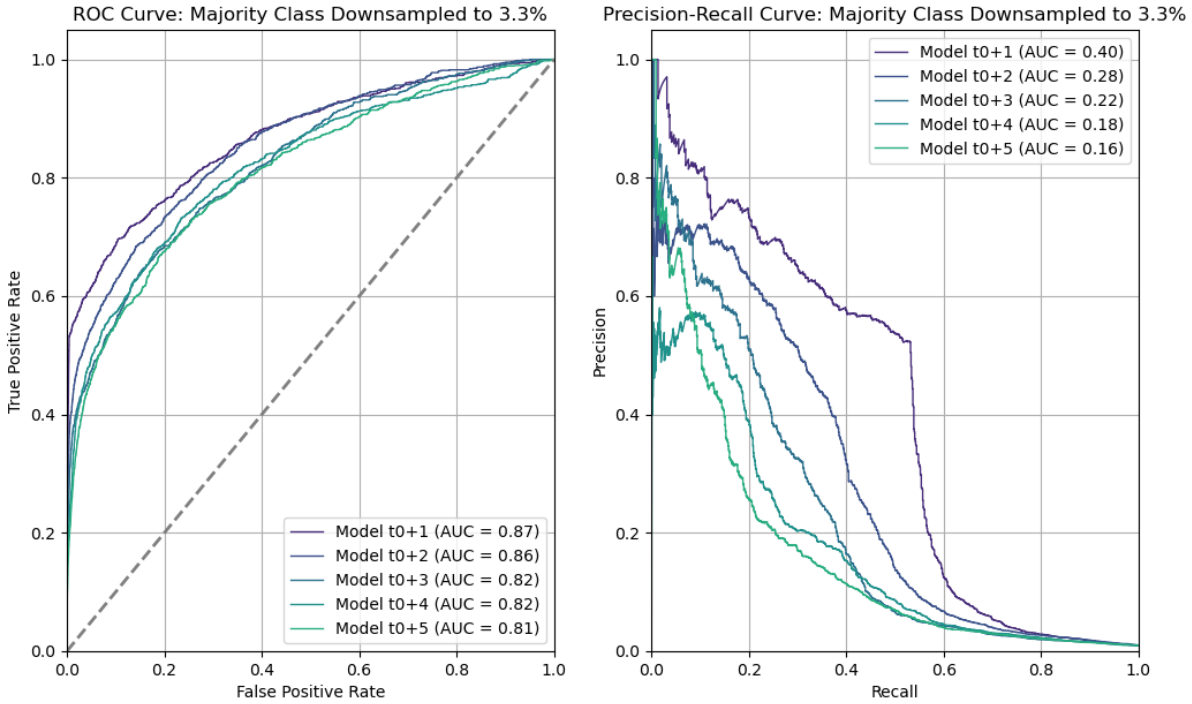



Figure 21: ROC and Precision-Recall plot of the temporal model trained on dataset D2. In Table 6 this includes Model 2-6.

Figure 21 presents the performance of the five temporal models trained on dataset D2, aimed at investigating potential improvements with a dataset

featuring a higher proportion of the majority class. While there is minimal difference between the AUC scores for these models compared to dataset D1, models $t_0 + 1$ and $t_0 + 2$ reveal slightly elevated values. The AUC-PR scores register marginal decreases across all models. Upon examining the precision, recall, and F1 values in Table 13, the models demonstrate similar behavior to those trained on dataset D1. Both groups of models achieve their most optimal results with a threshold of 0.4.

Table 13: Overview of precision, recall, and F1 values for the instantaneous model trained with dataset D2.

| **Model** | **Dataset** | **Threshold** | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|---|
| $t_0 + 1$ | D2 | 0.1 | 0.50 | 0.53 | 0.51 |
| $t_0 + 2$ | D2 | 0.1 | 0.41 | 0.37 | 0.39 |
| $t_0 + 3$ | D2 | 0.1 | 0.30 | 0.32 | 0.32 |
| $t_0 + 4$ | D2 | 0.1 | 0.20 | 0.32 | 0.25 |
| $t_0 + 5$ | D2 | 0.1 | 0.19 | 0.27 | 0.22 |
| $t_0 + 1$ | D2 | 0.4 | 0.57 | 0.44 | 0.50 |
| $t_0 + 2$ | D2 | 0.4 | 0.59 | 0.24 | 0.34 |
| $t_0 + 3$ | D2 | 0.4 | 0.60 | 0.14 | 0.23 |
| $t_0 + 4$ | D2 | 0.4 | 0.55 | 0.12 | 0.20 |
| $t_0 + 5$ | D2 | 0.4 | 0.46 | 0.11 | 0.18 |
| $t_0 + 1$ | D2 | 0.8 | 1.00 | 0.00 | 0.01 |
| $t_0 + 2$ | D2 | 0.8 | 0.69 | 0.01 | 0.02 |
| $t_0 + 3$ | D2 | 0.8 | 1.00 | 0.00 | 0.01 |
| $t_0 + 4$ | D2 | 0.8 | 0.67 | 0.00 | 0.00 |
| $t_0 + 5$ | D2 | 0.8 | 0.76 | 0.01 | 0.03 |

Table 14: Overview of the AUC and AUC-PR scores for all models for dataset D1 and D2.

| **Model** | **Dataset** | **AUC** | **AUC-PR** | **Brier** |
|---|---|---|---|---|
| $t_0 + 1$ | D1 | 0.86 | 0.42 | 0.0065 |
| $t_0 + 2$ | D1 | 0.85 | 0.30 | 0.0078 |
| $t_0 + 3$ | D1 | 0.84 | 0.25 | 0.0077 |
| $t_0 + 4$ | D1 | 0.83 | 0.22 | 0.0084 |
| $t_0 + 5$ | D1 | 0.80 | 0.16 | 0.0093 |
| $t_0 + 1$ | D2 | 0.87 | 0.40 | 0.0067 |
| $t_0 + 2$ | D2 | 0.86 | 0.28 | 0.0078 |
| $t_0 + 3$ | D2 | 0.82 | 0.22 | 0.0079 |
| $t_0 + 4$ | D2 | 0.82 | 0.18 | 0.0086 |
| $t_0 + 5$ | D2 | 0.81 | 0.16 | 0.0089 |

To complement model performance evaluation, we examine the calibration of the probabilistic classifiers using the calibration curve in Figure 22 and the Brier scores in Table 14.

A well-calibrated model would have points along the 45-degree diagonal line, represented by $x = y$. In predicting freezing weather occurrences, this alignment indicates that the model's estimated probabilities accurately reflect the actual likelihood of such events, mirroring the observed occurrence rates from the data.

For datasets D1 and D2, the Brier score is optimally low across all models, with D1 models showing slightly better performance. Observing the calibration plots, it becomes apparent that models trained on D1 exhibit smoother and less broken trends than those trained on D2. Notably, models $t_0 + 2$, $t_0 + 3$, and $t_0 + 4$ for D1 demonstrate tighter adherence to the perfectly calibrated line, particularly up to threshold values of 0.4 for both mean predicted probability and the fraction of positives. The plots for D2 models display more apparent changes, indicative of instances of boht under- and overconfidence.

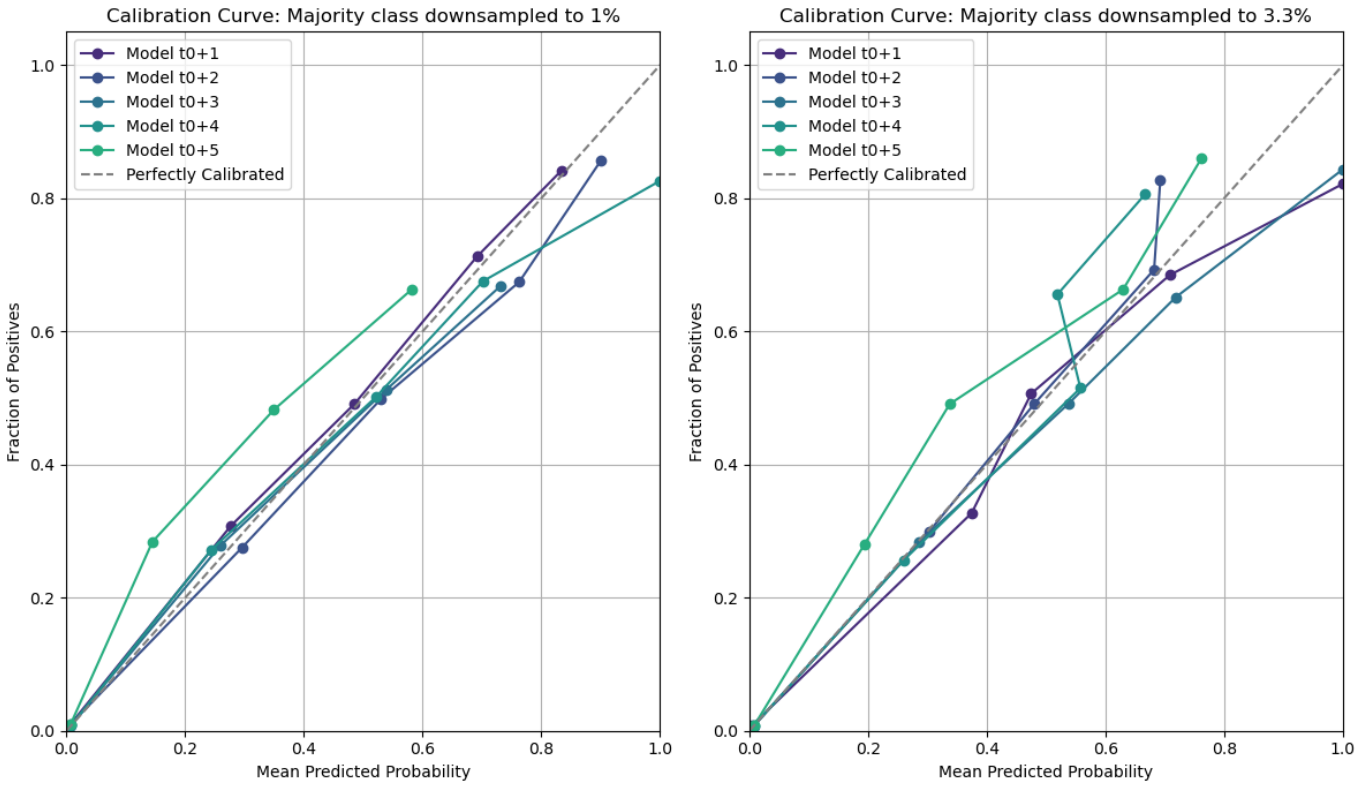



Figure 22: Calibration curves for all temporal models trained on dataset D1 and D2.

# 5 Discussion

The ML models developed and evaluated in this thesis aimed to explore the viability of predicting icing weather occurrences at airports. Leveraging input data from an NWP model and real-time observations at airports through registered METAR reports formed the basis of model training and testing. The overarching goal was to assess whether these classifiers could provide valuable and reliable probabilities, possibly supporting meteorologists in generating METAR reports.

This section will discuss and analyze the results of this process, emphasizing the promising insights obtained from the study. We will revisit the limitations outlined in the introduction, discussing potential challenges, potential areas for enhancement, and reflections on alternative approaches that

could have been pursued.

The expectations for the models were varied, as the dataset underwent significant downsampling of the majority class to achieve balance, thereby providing the model with a more robust foundation for generalization between negative and positive instances. Some positive side effects of having a small dataset are faster training, facilitating efficient hyperparameter tuning, and enabling exploration of various architectures and hyperparameter values [41]. It also prevents the model from getting too complex, where keeping the number of layers and units relatively low resulted in lower indications of overfitting and better performance on the validation set.

The results obtained from the temporal models demonstrate the efficacy of incorporating multi-timestep meteorological data development and METAR report observations. The most significant improvements are observed in the shorter time horizons, as more insecurity builds up as the timesteps move further away from $t_0$. Nevertheless, improvements are evident across all timesteps compared to the instantaneous and persistence models.

The difference in performance between the temporal models for datasets D1 and D2 is marginal, and determining which outcome is more optimal depends on the intended outcome. Considering the thesis objective, prioritizing the reduction of false positives emerges as a desirable goal. Precision takes precedence over recall in this scenario, emphasizing the importance of the model's ability to capture freezing weather instances confidently. This prioritization is particularly relevant for imbalanced datasets, where positive instances are scarce, necessitating accurate prediction and capture.

Conversely, minimizing false negatives is also in focus. Incorrectly classifying instances of icing weather as negative could lead to overlooking critical weather conditions. Thus, enhancing recall is necessary. It follows that there often is a trade-off between the two, where elevating the performance of one impacts the other. Another approach is to opt for a balance between the two.

Analysis of the data presented in Tables 12 and 13 reveals a delicate balance between precision and recall across all models and datasets. The most favorable outcomes are consistently achieved when observing the resulting values for a threshold of 0.4. Examination of dataset D1 at threshold 0.1 shows that recall values surpass precision. In contrast, for dataset D2, precision outperforms recall in models $t_0 + 1$ and $t_0 + 2$, with a slight decline observed in the following timesteps.

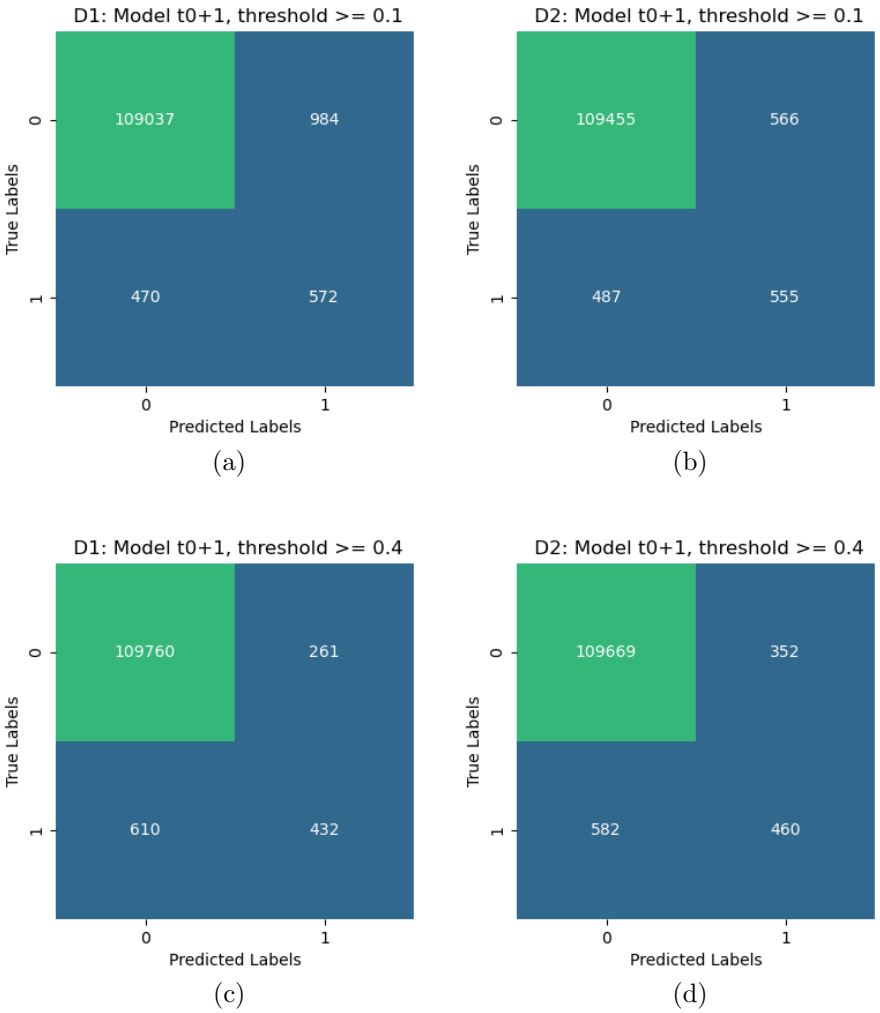



Figure 23: Confusion matrix for models $t_0 + 1$ and $t_0 + 2$ for datasets D1 and D2 at thresholds 0.1 and 0.4.

Figure 23 presents further insights into the difference between precision and recall for datasets D1 and D2, with the resulting confusion matrices for models $t_0 + 1$ and $t_0 + 2$ across both datasets at thresholds 0.1 and 0.4. For dataset D1 at threshold 0.1, depicted in Figure 23a recall values exceed precision, which is evident from the confusion matrix showing more false positives than true positives. Contrarily, false positives are markedly lower for the same model with dataset D2 at the same threshold, showcased in

Figure 23b.

When the threshold is increased to 0.4, both models exhibit a higher number of true positives over false positives, which is also reflected in the elevated precision values. Although the D2 dataset delivers a slightly higher number of false positives, it has fewer false negatives than D1.

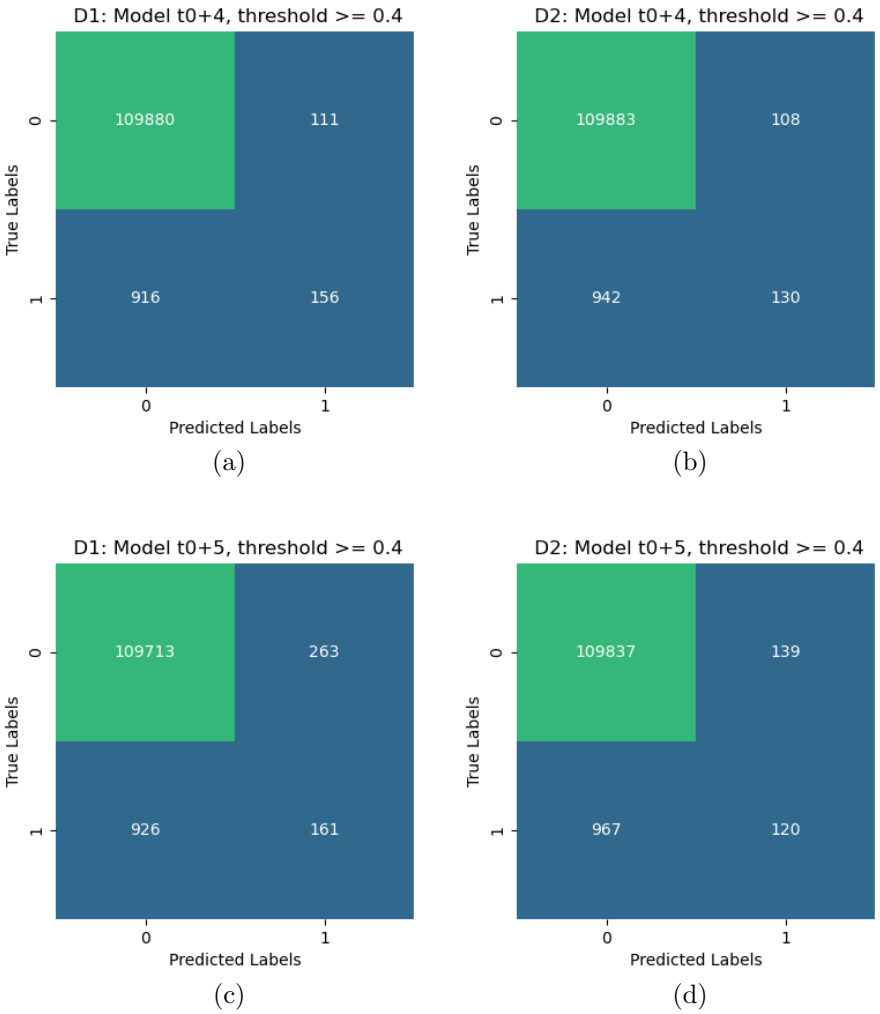



Figure 24: Confusion matrix for models $t_0 + 4$ and $t_0 + 5$ for datasets D1 and D2 at threshold 0.4.

As discussed earlier in this section, extending the prediction horizon with

additional timesteps introduces heightened uncertainty. With each increment in timesteps, there is a corresponding increase in false positives and false negatives, along with a decrease in true positives. Despite these challenges, the models still manage to capture positive instances. In Figure 24, the models $t_0 + 4$ and $t_0 + 5$ are represented with their respective confusion matrices for datasets D1 and D2 at a threshold of 0.4. Particularly for dataset D2 and model $t_0 + 4$ compared to $t_0 + 5$, Figure 24b and 24d, the transition from timestep four to five shows minimal fluctuations in prediction values, indicating the model's resilience in maintaining accuracy despite an expanded prediction horizon.

Revisiting the question about which dataset performs better after evaluating the results, the temporal models trained on dataset D1 provide marginally better performance. The overall difference in resulting values for AUC, AUC-PR, and Brier scores alongside the precision, recall, and F1 is slight but, as stated, slightly better for the D1 dataset.

The distinction between datasets D1 and D2 is the ratio of instances belonging to the majority class. Dataset D1 exhibits a more balanced distribution between the minority and majority classes, providing a slight advantage. This observation is consistent with the inherent challenges of class imbalance in ML and DL [41, 2]. Furthermore, the small size of the dataset aggravates performance limitations.

The skewed class proportion is one of the highlighted limitations of this study, introduced in Section 1.3.2. Downsampling the data, as performed in this study, leads to a considerable loss of information. Upsampling the data is an alternative approach to tackling the skewness in the data [27]. Such methods can involve duplicating existing data points or employing techniques such as the Synthetic Minority Over-sampling Technique (SMOTE), which generates new synthetic data points based on the existing ones [11]. Combining the downsampling of the majority class with the upsampling of the minority class could potentially enhance the model's performance by augmenting the training data.

The calibration curves depicted in 22, alongside the Brier scores detailed in Table 14, affirm the temporal models' reliability. A model's calibration refers to the alignment between its predicted probabilities and the actual outcomes. A well-calibrated model should accurately reflect the likelihood of events occurring. For instance, if the model predicts an 80% chance of icing weather in the next hour, the forecast is expected to match reality approximately 80% of the time. The results demonstrate a consistency be-

tween the models' predictions and observed outcomes, further indicating a trustworthiness of the produced information.

Another notable limitation highlighted in Section 1.3.2 is the AUTO-generated METAR reports. These reports are often inaccurate when forecasting icing weather occurrences, and the AUTO METAR reports containing FZ can almost be considered false positives. An approach to address this was to exclude these instances as a whole, which resulted in dataset D3, lower than half the size of D1. Because of that, in addition to the time constraint, it was only included in this thesis for training and testing in the persistence and instantaneous models for comparison. There were experiments with D3 and temporal models, but there was no evident or drastic change in the performance. An optimal approach here would be to upsample the data.

# 6 Conclusions

The objective of this thesis revolves around exploring the feasibility of employing ML techniques to predict icing weather events at airports. By leveraging input data derived from NWP models and real-time observations extracted from METAR reports, this study sought to assess the efficacy of the developed ML models and their potential implications for meteorology and aviation safety. Comprehensive evaluation and analysis have provided a better understanding and insight into the performance of these models, offering valuable and essential perspectives on their role in enhancing predictive capabilities in the field.

The motivation behind this work is to develop tools that can complement and validate meteorologists' efforts, particularly in the form of METAR reports. Integrating ML methodologies with human expertise aims to streamline this process, enhancing efficiency and reliability in weather forecasting endeavors.

The findings discussed in Section 5 demonstrate that the developed ML models exhibit promising abilities in forecasting icing weather events, with notable improvements observed in the temporal models' reliability and accuracy. Performing the training process with mainly two datasets, D1 and D2, where the primary difference was the size of the majority class, proposed the possibility of evaluating the effect of class imbalance and observed performance. Even with the resulting datasets being significantly smaller than the original and a loss of information through downsampling, the models demon-

strate their capability to deliver valuable and precise predictions, particularly evident when tested with a dataset that retained the original imbalance.

By integrating multi-timestep meteorological data and real-time observations in the input during training, these models delivered improved predictive capabilities compared to the persistence and instantaneous models. While there is an observed decrease in accuracy as the prediction horizon extends, it is essential to highlight that the models maintain the ability to accurately predict positive instances at four and five time steps ahead, as presented in Figure 23.

Building on the achievements of the study, several ideas for future research and potential approaches come into focus.

One potential approach, initially considered for inclusion in this thesis, was to explore whether incorporating meteorological data from a specified grid surrounding the airport could offer additional insights into the factors contributing to icing weather occurrences. Expanding the data coverage to encompass a larger geographical area, with a 10-30 km radius around the airport, would provide the model with more extensive training data. It could enhance the model's ability to make predictions by providing a stronger foundation to understand better the factors that lead to icing weather events.

Exploring alternative strategies to increase the volume of data is essential. These methods could involve extending the period from which data is collected or employing a combination of up- and downsampling techniques. By expanding the dataset, the models can discover hidden patterns more efficiently and gain more insights from the data. As the present study was data-driven by design and nature, future collaboration with domain experts in aviation meteorology can offer deeper insights and enhance the development and evaluation of the models. This could include incorporating the nuanced relationship between the different meteorological parameters and the processes leading to icing weather in the model design.

# References


[1] Kumar Abhishek et al. "Weather Forecasting Model using Artificial Neural Network". In: *Procedia Technology* 4 (2012). 2nd International Conference on Computer, Communication, Control and Information Technology( C3IT-2012) on February 25 - 26, 2012, pp. 311–318. ISSN: 2212-0173. DOI: https://doi.org/10.1016/j.protcy.2012.05.047. URL: https://www.sciencedirect.com/science/article/pii/S221201731200326X.

[2] Alhanoof Althnian et al. "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain". In: *Applied Sciences* 11.2 (2021). ISSN: 2076-3417. DOI: 10.3390/app11020796. URL: https://www.mdpi.com/2076-3417/11/2/796.

[3] Ulf Andrae et al. "A continuous EDA based ensemble in MetCoOp". In: (Jan. 2020).

[4] Alexandre M. Bayen and Timmy Siauw. "Chapter 14 - Interpolation". In: *An Introduction to MATLAB® Programming and Numerical Methods for Engineers*. Ed. by Alexandre M. Bayen and Timmy Siauw. Boston: Academic Press, 2015, pp. 211–223. ISBN: 978-0-12-420228-3. DOI: https://doi.org/10.1016/B978-0-12-420228-3.00014-2. URL: https://www.sciencedirect.com/science/article/pii/B9780124202283000142.

[5] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. "Evaluation Measures for Models Assessment over Imbalanced Data Sets". In: *Journal of Information Engineering and Applications* 3 (2013), pp. 27–38. URL: https://api.semanticscholar.org/CorpusID:52267786.

[6] Ben C. Bernstein. "Regional and Local Influences on Freezing Drizzle, Freezing Rain,and Ice Pellet Events". In: *Weather and Forecasting* 15.5 (2000), pp. 485–508. DOI: 10.1175/1520-0434(2000)015<0485:RALIOF>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434_2000_015_0485_raliof_2_0_co_2.xml.

[7] Yihua Cao, Wenyuan Tan, and Zhenlong Wu. "Aircraft icing: An ongoing threat to aviation safety". In: *Aerospace science and technology* 75 (2018), pp. 353–385.

[8] Yihua Cao, Wenyuan Tan, and Zhenlong Wu. "Aircraft icing: An ongoing threat to aviation safety". In: *Aerospace Science and Technology* 75 (2018), pp. 353–385. ISSN: 1270-9638. DOI: https://doi.org/10.1016/j.ast.2017.12.028. URL: https://www.sciencedirect.com/science/article/pii/S1270963817317601.

[9] Nora Casson et al. "Winter Weather Whiplash: Impacts of Meteorological Events Misaligned With Natural and Human Systems in Seasonally Snow-Covered Regions". In: *Earth's Future* 7 (Dec. 2019). DOI: 10.1029/2019EF001224.

[10] CBC News. *Investigators release new report into West Wind crash that killed 6 in Saskatchewan*. CBC News. 2024. URL: https://www.cbc.ca/news/canada/saskatoon/west-wind-crash-investigation-1.6228236 (visited on 04/30/2024).

[11] Nitesh V. Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *J. Artif. Int. Res.* 16.1 (June 2002), pp. 321–357. ISSN: 1076-9757.

[12] Ben Clarke et al. "Extreme weather impacts of climate change: an attribution perspective". In: *Environmental Research: Climate* 1.1 (June 2022), p. 012001. DOI: 10.1088/2752-5295/ac6e7d. URL: https://dx.doi.org/10.1088/2752-5295/ac6e7d.

[13] Jean Coiffier. "Half a century of numerical weather prediction". In: *Fundamentals of Numerical Weather Prediction*. Cambridge University Press, 2011, pp. 1–14.

[14] Timo Dimitriadis, Tilmann Gneiting, and Alexander I. Jordan. "Stable reliability diagrams for probabilistic classifiers". In: *Proceedings of the National Academy of Sciences of the United States of America* 118 (2020). URL: https://api.semanticscholar.org/CorpusID:231954837.

[15] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. "Activation functions in deep learning: A comprehensive survey and benchmark". In: *Neurocomputing* 503 (2022), pp. 92–108. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2022.06.111. URL: https://www.sciencedirect.com/science/article/pii/S0925231222008426.

[16] Wei Fang et al. "Survey on the Application of Deep Learning in Extreme Weather Prediction". In: *Atmosphere* 12.6 (2021). ISSN: 2073-4433. DOI: 10.3390/atmos12060661. URL: https://www.mdpi.com/2073-4433/12/6/661.

[17] Christopher A. T. Ferro. "Comparing Probabilistic Forecasting Systems with the Brier Score". In: *Weather and Forecasting* 22.5 (2007), pp. 1076–1088. DOI: 10.1175/WAF1034.1. URL: https://journals.ametsoc.org/view/journals/wefo/22/5/waf1034_1.xml.

[18] Richard Forbes et al. *Towards predicting high-impact freezing rain events*. eng. 2014 2014. DOI: 10.21957/xcauc5jf. URL: https://www.ecmwf.int/node/17334.

[19] Stefan Gössling et al. "Weather, climate change, and transport: a review". In: *Natural Hazards* 118 (June 2023), pp. 1–20. DOI: 10.1007/s11069-023-06054-2.

[20] Ismail Gultepe et al. "A Review of High Impact Weather for Aviation Meteorology". In: *Pure and Applied Geophysics* 176 (May 2019), pp. 1869–1921. DOI: 10.1007/s00024-019-02168-6.

[21] G. S. Handelman et al. "Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods". In: *AJR. American Journal of Roentgenology* 212.1 (2019), pp. 38–43. DOI: 10.2214/AJR.18.20224.

[22] C. Huang et al. "Performance Metrics for the Comparative Analysis of Clinical Risk Prediction Models Employing Machine Learning". In: *Circulation. Cardiovascular quality and outcomes* 14.10 (2021), e007526. DOI: 10.1161/CIRCOUTCOMES.120.007526.

[23] Anette Iital et al. *Estonia's Eighth National Communication Under the United Nations Framework Convention on Climate Change*. Estonian Environmental Research Centre and Ministry of the Environment, University of Tartu. Feb. 2023. URL: https://unfccc.int/sites/default/files/resource/Estonia%208th%20National%20Communication%20resubmission.pdf (visited on 05/03/2024).

[24] International Civil Aviation Organization. *Annex 3 to the Convention on International Civil Aviation: Meteorological Service for International Air Navigation*. Seventeenth. International Civil Aviation Or-

ganization, 2010. ISBN: 978-92-9231-507-8. URL: https://www.icao.int/airnavigation/IMP/Documents/Annex%203%20-%2075.pdf.

[25] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: (Feb. 2015).

[26] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[27] Pradeep Kumar et al. "Classification of Imbalanced Data:Review of Methods and Applications". In: *IOP Conference Series: Materials Science and Engineering* 1099.1 (Mar. 2021), p. 012077. DOI: 10.1088/1757-899X/1099/1/012077. URL: https://dx.doi.org/10.1088/1757-899X/1099/1/012077.

[28] Amy McGovern et al. "Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather". In: *Bulletin of the American Meteorological Society* 98.10 (2017), pp. 2073–2090. DOI: https://doi.org/10.1175/BAMS-D-16-0123.1. URL: https://journals.ametsoc.org/view/journals/bams/98/10/bams-d-16-0123.1.xml.

[29] Francis Nahm. "Receiver operating characteristic curve: overview and practical use for clinicians". In: *Korean Journal of Anesthesiology* 75 (Jan. 2022). DOI: 10.4097/kja.21209.

[30] Meera Narvekar and Priyanca Fargose. "Daily weather forecasting using artificial neural network". In: (2015). URL: https://research.ijcaonline.org/volume121/number22/pxc3905088.pdf.

[31] Norwegian Meteorological Institute (MET Norway). *MetCoOp Project*. https://www.met.no/en/projects/metcoop. Year. (Visited on 05/03/2024).

[32] Gilles Notton and Cyril Voyant. "Chapter 3 - Forecasting of Intermittent Solar Energy Resource". In: *Advances in Renewable Energies and Power Technologies*. Ed. by Imene Yahyaoui. Elsevier, 2018, pp. 77–114. ISBN: 978-0-12-812959-3. DOI: https://doi.org/10.1016/B978-0-12-812959-3.00003-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128129593000034.

[33] Ismoilov Nusrat and Sung-Bong Jang. "A Comparison of Regularization Techniques in Deep Neural Networks". In: *Symmetry* 10.11 (2018). ISSN: 2073-8994. DOI: 10.3390/sym10110648. URL: https://www.mdpi.com/2073-8994/10/11/648.

[34] Sanjay Mathur Paras, Avinash Kumar, and Mahesh Chandra. "A feature based neural network model for weather forecasting". In: *International Journal of Computational Intelligence* 4.3 (2009), pp. 209–216. URL: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b16997cefef79d1768a274bdf0201aac94a717ef.

[35] Marius Paulescu, Eugenia Paulescu, and Viorel Badescu. "Chapter 9 - Nowcasting solar irradiance for effective solar power plants operation and smart grid management". In: *Predictive Modelling for Energy Management and Power Systems Engineering*. Ed. by Ravinesh Deo, Pijush Samui, and Sanjiban Sekhar Roy. Elsevier, 2021, pp. 249–270. ISBN: 978-0-12-817772-3. DOI: https://doi.org/10.1016/B978-0-12-817772-3.00009-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128177723000094.

[36] Lutz Prechelt. "Early Stopping - But When?" In: *Neural Networks: Tricks of the Trade*. Ed. by Genevieve B. Orr and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 55–69. ISBN: 978-3-540-49430-0. DOI: 10.1007/3-540-49430-8_3. URL: https://doi.org/10.1007/3-540-49430-8_3.

[37] Zhiyong Pu and Eugenia Kalnay. "Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation". In: *Handbook of Hydrometeorological Ensemble Forecasting*. Ed. by Qiuhong Duan et al. https://www.inscc.utah.edu/~pu/6500_sp12/Pu-Kalnay2018_NWP_basics.pdf. Berlin, Heidelberg: Springer, 2018.

[38] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, Third Edition*. Third. Packt Publishing, 2019. ISBN: 978-1-78995-575-0.

[39] *Receiver Operating Characteristic (ROC) curve*. URL: https://commons.wikimedia.org/w/index.php?curid=109730045 (visited on 04/29/2024).

[40] Yoram Reich and S.V. Barai. "Evaluating machine learning models for engineering problems". In: *Artificial Intelligence in Engineering* 13.3 (1999), pp. 257–272. ISSN: 0954-1810. DOI: https://doi.org/10.1016/S0954-1810(98)00021-1. URL: https://www.sciencedirect.com/science/article/pii/S0954181098000211.

[41] Anastasiia Safonova et al. "Ten deep learning techniques to address small data problems with remote sensing". In: *International Journal of Applied Earth Observation and Geoinformation* 125 (2023), p. 103569. ISSN: 1569-8432. DOI: https://doi.org/10.1016/j.jag.2023.103569. URL: https://www.sciencedirect.com/science/article/pii/S156984322300393X.

[42] Shibani Santurkar et al. "How Does Batch Normalization Help Optimization?" In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/905056c1ac1dad141560467e0a99e1cf-Paper.pdf.

[43] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. "Activation functions in neural networks". In: *Towards Data Sci* 6.12 (2017), pp. 310–316.

[44] Jana Sillmann et al. "Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities". In: *Weather and Climate Extremes* 18 (2017), pp. 65–74. ISSN: 2212-0947. DOI: https://doi.org/10.1016/j.wace.2017.10.003. URL: https://www.sciencedirect.com/science/article/pii/S2212094717300440.

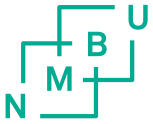
[45] Simple Flying. *The Tragic Story Of Iran Air Flight 277*. 2022. URL: https://simpleflying.com/iran-air-flight-277-story/ (visited on 04/30/2024).

[46] Yingjie Tian and Yuqi Zhang. "A comprehensive survey on regularization strategies in machine learning". In: *Information Fusion* 80 (2022), pp. 146–166. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2021.11.005. URL: https://www.sciencedirect.com/science/article/pii/S156625352100230X.

[47] Peter A G Watson. "Machine learning applications for weather and climate need greater focus on extremes". In: *Environmental Research Letters* 17.11 (Nov. 2022), p. 111004. DOI: 10.1088/1748-9326/ac9d4e. URL: https://dx.doi.org/10.1088/1748-9326/ac9d4e.

Figures of neural networks adapted from `https://tikz.net/neural_networks/`.

Figure 1: Figure created with dummy METAR data, selfmade with LucidChart.




**Norges miljø- og biovitenskapelige universitet**
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway