**University of South Bohemia in České Budějovice**
**Technische Hochschule Deggendorf**
**Fakultät Angewandte Informatik**

Studiengang Master Artificial Intelligence and Data Science

# Deep Learning zur Brunsterkennung bei Milchkühen – Zeitreihenmodellierung mit LSTM-Netzwerken

# Deep Learning for Estrus Detection in Dairy Cattle – Temporal Modeling Using LSTM Networks

Masterarbeit zur Erlangung des akademischen Grades:

*Master of Science (M.Sc.)*

an der Technischen Hochschule Deggendorf

Vorgelegt von:
Muhammed Mufad Ambhalathuveetil
Matrikelnummer: 12301340

Am: 29. Januar 2026

Erstbetreuer:
Markus Eider, M.Sc.

Betreuer aus dem Unternehmen:
Michal Bechny, PhD.

# Bibliographic Information, Annotation, and Declaration

Ambhalathuveetil, Muhammed Mufad, 2026: Deep Learning for Estrus Detection in Dairy Cattle – Temporal Modeling Using LSTM Networks. Master's Thesis, in English– 67p. Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic. Faculty of Applied Computer Science, Deggendorf Institute of Technology, Deggendorf, Germany.

## Annotation

The thesis focuses on the application of artificial intelligence method for estrus detection in dairy cows by using time series data from wearable sensor devices. A Long Short-Term Memory (LSTM) based modelling approach is applied to understand behavioural patterns associated with estrus cycle and is evaluated through comparative analysis with baseline models. The research aims to improve reproductive decision-making in the field of precision livestock farming.

## Declaration

I declare that I am the author of this qualification thesis and that in writing it I have used the sources and literature displayed in the list of used sources only. I further declare that I have used ChatGPT, generative artificial intelligence tool(s) or service(s) in accordance with academic ethics for the purpose of language correction, sentence clarity, and LaTeX structure.

Place, Date:
České Budějovice, January 29, 2026
Signature:

Muhammed Mufad Ambhalathuveetil

# Erklärung

Name des Studierenden: Muhammed Mufad Ambhalathuveetil

Name des Betreuenden: Markus Eider, M.Sc.

Thema der Abschlussarbeit:

Deep Learning zur Brunsterkennung bei Milchkühen – Zeitreihenmodellierung mit LSTM-Netzwerken

1. Ich erkläre hiermit, dass ich die Abschlussarbeit gemäß § 35 Abs. 7 RaPO (Rahmenprüfungsordnung für die Fachhochschulen in Bayern, BayRS 2210-4-1-4-1-WFK) selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

2. Diese Masterarbeit wurde in Zusammenarbeit mit dem Unternehmen Farmtec a.s. durchgeführt. Einige Informationen über das interne Wissen des Unternehmens oder bestimmte von ihm verwendete Technologien mussten aufgrund der unterzeichneten Vertraulichkeitsvereinbarung in dieser Arbeit weggelassen werden.

Deggendorf, 29.01.2026 .....
Datum

.....
Unterschrift des Studierenden

# Acknowledgement

This thesis is a display of my academic learning combined with the real-world industrial problem. The completion of this thesis has been made possible through the help, support, and encouragement of individuals as well as institutions, to whom I sincerely express my gratitude.

First and foremost, I would like to thank Mr. Michal Bechny for his support throughout this project. His willingness to explore new approaches, encouragement to pursue improvement, and constructive feedback shaped this work. I am also grateful to Mr. Marek Vrhel for providing strong industry knowledge and practical insights into estrus detection, which grounded this study in real-world challenges. I would also like to extend my sincere thanks to Mr. Markus Eider for his academic guidance, insightful feedback, and practical suggestions, which have significantly improved the thesis's clarity, organization, and scientific content.

Finally, I gratefully acknowledge Farmtec a.s. for their collaboration in experimental design, the provision and processing of industrial data, and the provision of technical expertise. The availability of high-quality real-world data enabled a realistic evaluation of the presented techniques and was critical to the study's success.

# Statement On The Use Of Digital and Artificial Intelligence (AI) Tools

The thesis has been constructed independently, and digital and AI based tools are used solely for support purposes, rather than as a replacement for original work and scientific reasoning.

ChatGPT was used for language correction, sentence clarity, and LaTeX structure. All suggestions made by it were reviewed and revised. Grammarly was used to find grammatical and styling issues. References and citations were managed using Zotero. The correctness of all cited literature was manually verified for quality.

AI and digital tools were used as aids to improve efficiency and presentation, while all scientific inputs and final decisions were made independently at all points.

# Abstract

Detecting the fertility period in cattle is critical at cattle farms since it has direct influence to the herd productivity, efficient milk production, and overall profitability. That is where the role of Estrus Detection (ED) becomes central. Although there are many manual techniques, termed as traditional methods, applied to detect estrus on time, there is a significant chance of minor errors in determining the accurate time of estrus, which may induce economic inefficiency and gradually cause drastic changes in overall farm production.

The traditional methods, including visual observation and hormonal change calculation, have the possibility of miscalculations in variations due to minor human errors and machine errors, which need to be avoided. The lack of sensitivity, high cost, and time-consuming processes induced the potential of AI to be included in the process of ED.

This thesis examines the application of AI in the field of dairy farming by applying Long Short-Term Memory (LSTM) networks, to time-series data for capturing the behavioural patterns indicating estrus. The proposed approach is evaluated through comparison with baseline models.

The results demonstrate that the LSTM approach effectively captures estrus patterns, while achieving comparable performance with the classical baseline methods, with reduced false alarms. The findings highlight the efficiency of the model in capturing estrus events and showcases its capability to support improved reproductive decision making in diary farming. This opens possibilities to enhance precision livestock farming by advancing its potential through Deep Learning (DL).

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

**AI** Artificial Intelligence

**AUC** Area Under The Curve

**CNN** Convolutional Neural Network

**DL** Deep Learning

**ED** Estrus Detection

**IoT** Internet of Things

**IQR** Interquartile Range

**LR** Logistic Regression

**LSTM** Long Short-Term Memory

**MC** Monte Carlo

**ML** Machine Learning

**PR** Precision-Recall

**RF** Random Forest

**RNN** Recurrent Neural Network

**ROC** Receiver Operating Characteristics

**SVM** Support Vector Machine

**TS** Time-Series

# 1 Introduction

The rise in the demand for dairy products fueled the growth of the dairy Industry. Milk production is the most driving part of the Industry. The quality and quantity of it determine the profitability and growth of the farm [1]. If Milk production is the pillar of the Industry, the dairy herd reproduction is the foundation of this pillar, so too is the dairy Industry. It maintains production continuity and increases the number of cattle in the Industry. Along with enhancing the high-production stage, it lowers veterinary costs and reduces the duration of the low-production phase.

The global expansion in demand for dairy products boosted growth in the dairy herd management market. The continuous population growth, which led to high per capita consumption of milk products, shifted the global dairy market. To increase the profitability, farms are now utilising technologies, including AI, to manage herd productivity [2] [3]. The rise in herd productivity called into question the efficiency of traditional methods and underscored the need to adopt technological solutions.

The Milk production is initiated after the calving stage of the dairy cattle. It is where the estrus cycle comes to the center. The fertility period is termed as estrus, which is also referred to as 'heat', during which the cow is sexually receptive and exhibits behavioural changes that indicate her readiness to be bred [4]. Based on the internal information provided by Farmtec a.s. estrus occurs in heifers at approximately 8 months of age, although insemination is suitable from only 14 months of age. Estrus commonly occurs 40 to 42 days after calving, farmers start looking for estrus again and proceed with the insemination depending on the reproductive management [5]. The estrus cycle lasts 21 days in cattle, which consists of four stages: proestrus, estrus, metestrus, and diestrus. Figure 1.1 demonstrates the duration of each stage in the estrus cycle. The estrus phase lasts only 6–30 hours, which is the time when the cattle exhibit mating desire. The production of milk starts after the calving stage, and the milk collection should be paused during the dry period before the next calving stage [6].

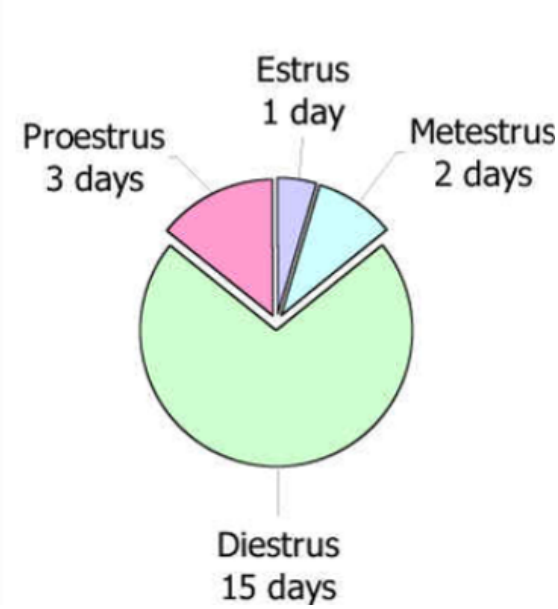


Figure 1.1: Estrus Cycle Stages(from [7])

Early detection of estrus helps maintain production continuity and increases the number of cattle in the Industry. While Traditional detection methods, such as visual observations, can delay detection, and Hormone assays are often more expensive. The sensor and AI-based monitoring enabled more systematic ED compared to traditional methods [8]. The features of cattle movements were tracked using a wearable sensor and captured as detailed Time-Series (TS) data.

This helps to address the backlogs that persisted in manual detection through continuos monitoring and effective data driven decision support. In this context, DL approaches such as LSTM may provide incremental improvements for ED as they as suitable for TS data. Within the scope of the thesis, the LSTM framework will be employed to represent temporal dependencies and to enable the reliability analysis, which involves the estimation of uncertainty.

## 1.1 Motivation

The dairy Industry focuses on maintaining herd productivity through efficient early ED. Each missed estrus cycle results in significant economic losses for the company. A missed heat in dairy cows would cost the company around 50-60 euros per cow [9]. According to communication with Farmtec a.s., the current estimated cost of missed estrus has increased in recent years, and it is approximately 83 euros per cow over 21 days, while it depends on the cost of the feed. Failing to perform insemination at the right time leads to a low pregnancy rate, resulting in the prolongation of the low-production phase. In addition, because of the absence of milk production and consuming the same amount of feed as in the peak lactation period, there is a chance of problems with excessive fat accumulation in the cow's body. This leads to serious issues after calving, such as difficulties in starting the new lactation, impaired reproduction, and

negative energy balance [10] [11]. Labour-intensive and expensive manual detection failures highlight the need for modern, efficient solutions.

The development of new technologies has compelled farmers to adopt them to enhance operational efficiency and improve herd management. The Internet of Things (IoT) and data-driven insights enabled by sensor networks allow the continuous monitoring of physiological and behavioural patterns linked to the estrus cycle [12]. The sensor networks help to track down the behavioural and physical changes that occur in cows using a wearable device such as vitalimeter. It is a wearable neck device that monitors the physiological and biological factors, which are continuously monitored on an hourly basis to produce TS data. As per the information provided by Farmtec a.s., the main principle behind it is the evaluation of data from a 3 axis gyroscope which measures the sensitive vibrations and interprets the behaviour based on this. This continuous tracking with the help of a wearable device helps to collect data of the cattle, which is then stored in a cloud and used for practical data analysis to find biological patterns. Farmtec a.s. uses this technology to detect health problems, feed intake problems, and analyse estrus patterns. Figure 1.2 shows the innovative IoT-based health monitoring system designed for dairy cows. Using this sensor technology, the collected data can be used to identify patterns for early heat detection effectively. The patterns of cows exhibiting the symptoms of estrus can be captured using this technology, which can improve the decision making in the farm sector.

Additionally, the tracking of behavioural, physiological, and biological factors of dairy cows ensures the development of precision livestock farming. At the same time, it assures that the reproductive health and living environment of the cattle are maintained at the safest level [13]. It helps to improve the lifespan of the cattle by detecting any health issues early. The integration of IoT with farm technology opens a vast world of innovations that can be implemented in the agricultural sector. Innovative farming technology, combined with precision livestock farming, is going to revolutionise the dairy industry.

This technical advancement, along with the challenge of analysing complex temporal patterns in dairy cattle activity, highlights the suitability of LSTM networks. Their ability to specifically remember or forget information makes them suitable for capturing extended temporal patterns, which makes them a better tool in the field of ED [14].

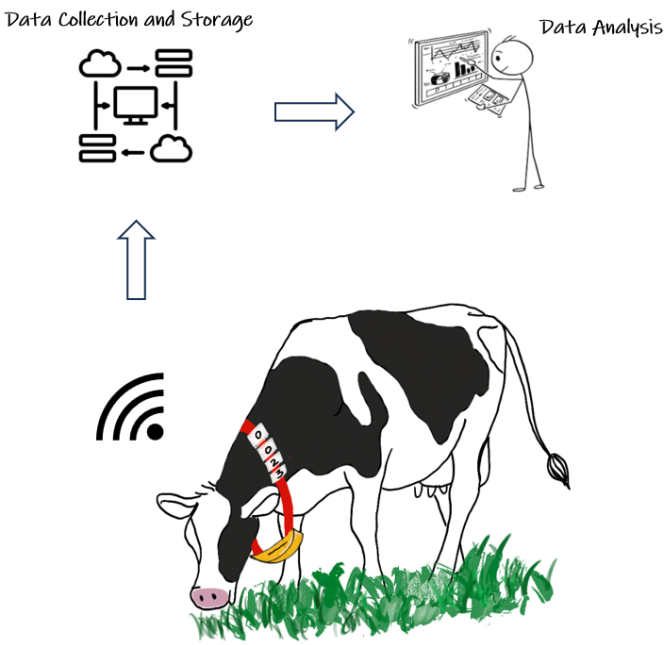


Figure 1.2: Data Collection illustration with wearable device adapted from [15]

## 1.2 Problem Statement

Dairy cows exhibit different behavioural changes during the estrus period, such as increased activity and mounting behaviour, which makes ED difficult. The shorter heat duration also makes the detection complicated to find the optimum breeding window [16]. This factor also risks the efficiency in manual ED. Along with the above factors, the increase in herd population, inconsistency in behavioural changes, absence of visible signs, and external factors make the visual detection inefficient.

A few of the automated heat detection methods have been found to struggle because of false estrus signals, which mimic estrus-related activity. The false-positive alerts produced by these methods have called their quality into question [17]. Although the strategies using AI incorporating Machine Learning (ML) have been overcoming all these issues, the class imbalance, which referred as the uneven distribution of positive and negative classes in the TS data captured by the sensors, has been a significant challenge to the existing models. The estrus events are rare in the whole TS data, with estrus events represents only 0.81% and non-estrus events accounts for remaining 99.19%.

The lack of extensive, real-world datasets for data validation is a significant setback in the field. In addition, the false alarms reduced the authenticity of the AI strategy among the farmers [18]. This enhances the need for a new approach that captures the TS dynamics captured

from real-world data with fewer false alarms. As each false alarm points to an incorrect insemination, it will be an economic burden to the industry.

Although there is significant development in the field of integrating innovative technologies in dairy farming, There is still a need to improve the applicability of these technologies in practical, real-world scenarios. The variations in climatic conditions affecting the sensor functioning, the need for extensive storage resources to store large amounts of data, and the variations in feeding patterns often make the applicability challenging. The individuality among cows, especially in terms of pregnancy rates and ages, challenges the generalisation capability of the patterns [19].

The approaches exploited needs to be improved in terms of accuracy, generalisability, and applicability. The challenges hidden behind data imbalances, highly complex data storage, and false alarms need to be addressed efficiently; only then can the Industry be presumed to be advanced in ED [9].

## 1.3 Research Questions

The thesis focuses on the project objectives and experimental collaborative work with Farmtec a.s. It aims to answer the following questions for effective early ED:

1. How accurately does the LSTM network model generalise to unseen dairy cows under practical conditions using real-world data?

2. How does the length of the prediction window influence the model's predictive performance, and how can its practical applicability be improved through different strategies?

Practical applicability refers to the model's ability to perform precisely in real-world farm environments, with greater generalisation and adaptability.

The questions were articulated based on the challenges discussed in the Motivation and Problem Statement sections. The inefficiency of traditional methods and the false alarms produced by automated methods indicate the need for an accurate AI-based model to predict estrus signals. Hence, the questions focus on how precisely the LSTM model can generalise the estrus signals and apply them under real-world conditions.

# 2 Domain Overview

## 2.1 Understanding Estrus Cycle

To enhance milk production, it is necessary to build a larger herd on the farm. Each herd on the farm contributes to the company's higher milk production. For a cow to maintain continuous milk production, it must calve regularly. The hormonal changes during the period of calving make the cow produce milk to feed their calves. The period of milk production is termed the lactation curve. The production increases progressively to a peak production point and then decreases to an off-stage termed as the drying period. This stage prepares the cow for her next lactation cycle [20]. To continue the lactation cycle, cows must calve, and the fertility period should be identified in time.

The early detection of estrus is important to ensure that the insemination takes place at the right time Moreover, it should coincide with the ovulation process to ensure pregnancy. Although ED has many challenges due to the short duration of estrus and subtle behavioural changes, the estrus cycle of the cow is defined as the time period from one event of estrus to the next. The cycle averages 21 days and is classified into four periods, including estrus, metestrus, diestrus, and proestrus [6] [21].

The estrus period is the stage that indicates readiness for ovulation and sexual receptivity. This is the period when cows display various behavioural and physical signs, including the stand-to-be-mounted behaviour, increased activity, restlessness, and vocalisation. These signs vary among cows, increasing variability. The period usually lasts 6 to 30 hours, With an average of 18 hours, it makes it difficult to detect early. The next stage occurs during metestrus, which lasts 3 to 5 days, followed by diestrus, which lasts 12 days. The estrus period enhances hormonal fluctuations, and the cow is no longer receptive to mating. The proestrus stage is the beginning phase of estrus, which leads to ovulation [6]. Figure 1.1, presented in the introduction, demonstrates each stage of the estrus cycle along with its duration.

The shorter duration of estrus makes it challenging to detect in time. Early detection enhances the optimal reproductive capability in cows, and hence maintains the lactation cycle. The herd population growth is also linked with the estrus cycle. A missed estrus delays the calving time, hence affects the lactation cycle and decreases milk production. The economy of the dairy industry completely relies on the accurate timing of ED [22]. It defines the biological and economic requirement, which is essential for the growth of the dairy industry. Figure 2.1 shows the overview of the reproductive cycle along with the sequential phases of the dairy cow. It visualises the progression from proestrus to diestrus, then to pregnancy, progressing through calving, lactation, and the dry period.

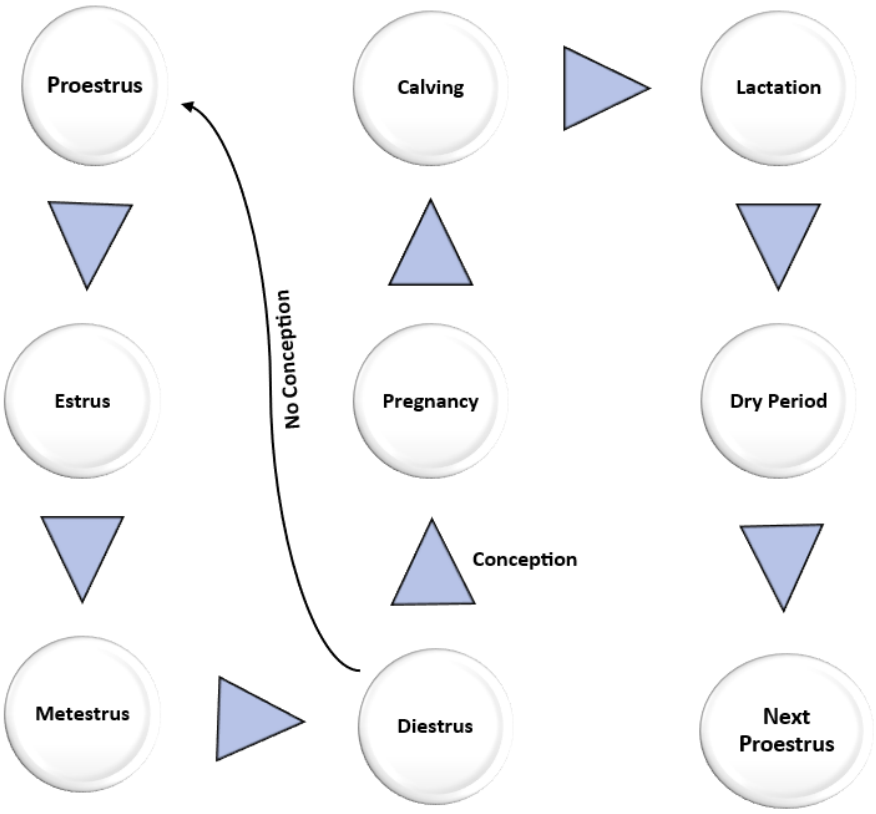


Figure 2.1: Overview of cow reproductive cycle(from [6])

Small-scale farms mostly rely on manual observations, and hence they face significant challenges that require more labor and a heavy workload to observe the cow on a 24-hour basis. The higher herd population makes manual observation of estrus very hard. The missed rate of estrus is high in the case of manual observation, and this leads to false breeding or open days and hence, it will cause economic burden to the farms. The term open days refers to the continued non-pregnant days in cows [23].

The automated ED methods using sensor data and AI may help to minimise the challenges of heavy workload and reduce human interaction. Detecting estrus on time not only increases milk production but also ensures healthier cows by reducing open days and ensuring the productive life cycle. However, the shorter heat duration, variability in cow behavioural signs, increased herd population, and increased workload have fueled the introduction of sensor-based monitoring in cows.

These methods are now transforming into data-driven processes. This not only helps to monitor the physiological conditions and biological factors that help in maintaining healthy feeding habits, but also assists in planning proper breeding treatments in cows. The following section discusses the evolution of traditional methods to sensor-based systems such as the vitalimeter, which tracks the cow on a 24-hour basis to monitor their biological and physiological patterns to provide reliable insights for reproduction.

## 2.2 Sensor Based ED

The introduction of sensor-based tracking of cows reduced the complexity of the ED process. It transformed the whole Industry in terms of herd and reproductive management. The significance of the sensor came to the scene when there were a lot of misjudgments occurring due to manual heat detection methods. Research shows that during the first 60 days after pregnancy, 40% of cows exhibit silent estrus signals, which are very hard to capture [8]. All the kinds of limitations possessed by traditional manual detection methods, along with the difficulty in managing larger herds and the extensive labour requirements, highlighted the growth of sensor technology in the dairy industry.

A wide variety of sensor technologies has been implemented in Industry based on the purposes for which they are required. Some focus entirely on health tracking, while others are attached to collars or legs, which have different functionalities that provide deeper insights into behavioural and physiological data captured through the sensors [8]. Sensors are also used for different purposes to ensure continuous health monitoring. The reproductive status of cows.

Among the highly efficient and innovative sensors, Farmtec a.s. Developed a continuous monitoring wearable sensor, vitalimeter. The vitalimeter plays an important role in this study by providing TS data through the continuous capture of hourly behavioural and physiological variations. By regularly capturing the physical activity, eating, and rumination patterns together with assessing the production and reproductive statistics of dairy cattle, it gives a better overview to the farmers. The vitalimeter collects the hourly data of each cow individually and processes this data with the help of a motion activity antenna. The motion activity antenna aggregates the data from multiple cows and transfers it to the online application. The processed data are visualised in an online application, and all the data can be accessed through this platform. This ensures that the data of each herd is always available. It mainly focuses on detecting the deviations from standard patterns that occur during the estrus period.

The vitalimeter is designed as a wearable collar type, which ensures wearability and accessibility for each cow. The device reliably detects the first signs of heat because of this design. The device collects high-resolution and TS data, which gives a clear idea about the cow's behavioural changes during estrus events. By capturing data hourly, it ensures that there are no missed variations [24] [25]. Figure 2.2 shows the workflow of the vitalimeter from data collection to data storage.

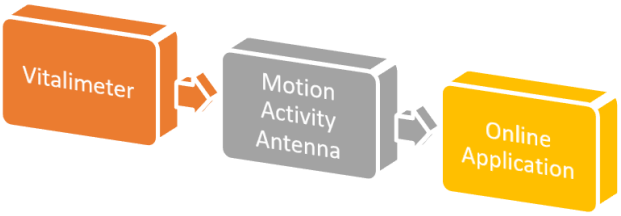


Figure 2.2: Data Collection using vitalimeter from [25]

Each dairy cow is given a specific identification number by the system, and individual cow data are stored. The stored TS data can be accessed and monitored through the central database system. This raw data has to be collected and analysed for accurate ED. The data are preprocessed to remove noise and should be analysed in depth using DL models.

## 2.3 Traditional vs Sensor-Based Detection

The main difference between traditional and sensor-based detection lies in the method used to analyse the behavioural, physiological changes, and how those analyses can be assessed for better decision-making. Table 2.1 indicates the different processes carried out in dairy farms through traditional methods and sensor-based automated systems, along with their key differences. It demonstrates the significant differences and the contrast between the methods in the livestock management process. It showcases the ineffective traditional practices, which create challenges that affect the health and productivity of dairy cattle.

Table 2.1: Comparison between Traditional and Sensor-Based ED Methods

| Processes | Traditional Methods | Sensor-Based Automated Methods |
|---|---|---|
| **Data Collection** | Visual observation of cows for visible signs such as increased activity or mounting tendency [6]. | Uninteruppted monitoring using sensors for capturing TS data[12]. |
| **Detection reasoning** | Based on the observed behavioural and physiological signs and farmer experience [9]. | Relies on detailed analysis of the TS data, analysing all factors. |
| **Accuracy** | Chances of human errors, Subjective to the farmer. | Higher accuracy due to the advanced data analysis and pattern matching [26]. |
| **Monitoring Duration** | Difficult to observe 24 hours. | Continuous 24 hour observation [12]. |
| **Labour Requirement** | Heavy requirement of labour to maintain the observation and keep the data [9]. | Less labour requirement due to an automated environment. |
| **Data storage** | Paper-based systems [9]. | Real-time monitoring dashboards and cloud-based storage. |
| **Cost and Maintenance** | Low initial cost and higher maintenance cost [9]. | Higher initial cost and lower maintenance cost [27]. |
| **Reproductive rate** | Low pregnancy rate due to missed detection [9]. | Improved Pregnancy rates through accurate detection [8]. |

The traditional method, including Visual Observation, detects estrus through visible signs such as increased movement and mounting behaviour [28] [29]. Although this method is considered accurate during the estrus period, the labour-based visualization of signs in larger herds creates difficulties, and the shortage of labour reduces the accuracy of detections. The detection rate is often below 50%. The major challenge during the estrus period is not only to detect the events on time but also to ensure that no events are missed. Each missed estrus event causes significant loss to the Industry and also affects the reproductive health of the cow. The physical workload, combined with the requirement of 24-hour observation, makes manual detection highly labour-intensive and inefficient for large farms [29].

The above-mentioned problems and limitations of traditional methods reduce detection accuracy and highlight the need for a data-driven process to overcome these challenges. Sensor technologies have been implemented in the cattle farming industry to address the above limitations. They play an essential role in modern dairy farming.

Along with the fact that they emphasise real-time monitoring of each individual herd in the farm, they also provide sequential temporal data for further analysis. Moreover, by reducing the human errors in traditional methods, they pave the path for the widespread growth of precise ED.

## 2.4 Challenges in Automated ED

Although the introduction of sensor-based monitoring technology has revolutionised the Industry by improving efficiency and accuracy in ED, multiple challenges also make it pull backwards, which doubts its reliability and practical appropriateness. Since the sensor-based systems monitor individual cows, the challenge mainly arises because of the individual variability in physical and biological conditions. The extreme environmental conditions affecting sensor, scalable deployment, and the areas with limited energy supply, including power and internet, also make the sensor-based systems difficult to manage [8].

Additionally, Farmtec a.s. has been faced multiple challenges such as signal interference that was not present during the initial installation setup, wrong animal registration, the need of long battery life and the sensor attachment difficulty with cows. To overcome this, there must be a need for periodic maintenance of the sensor devices, and it showcases the need for more labour in the farm [27]. There must be a need for backup plans to replace the malfunctioning ones, since each hour of tracking is necessary in the field to ensure accurate reproductive health. Although the sensors have been placed with specific design and suitability according to the cattle, such as neck collar design, this sometimes leads to damage easily because of the cattle's behaviour or intervention [8]. This makes interruptions in the data collection, and needs to be addressed early. The periodic maintenance requires the continuous replacement of the battery, as well as continuous health tracking, to ensure the proper functioning of the sensors.

One of the significant obstacles lies in the reliability of the tracked outcomes because of false positives and false negatives. The inaccurate monitoring or the false signals produced by the cattle are the main reasons behind this. The increased activity may be because of different reasons, such as regrouping or feeding requirements. This can be misinterpreted and creates false signals [27]. Along with this, another major issue lies in the data imbalances in the dataset. The difficulty in detecting the heat arises from the fact that the time span of estrus is very short. It only occurs for a few hours in a 21-day estrus cycle. This itself indicates that most of the data collected should be from non-estrus periods rather than estrus. The unevenly skewed dataset makes the detection hard since the negative class overtops the positive signals.

Addressing this data imbalance requires a highly precise and highly intelligent DL framework. Lastly, the need for improving the generalisation also arises as a major challenge in the field. The practical applicability in real farm conditions is only applicable in this way. This challenge arises the need for a highly distinguishable framework that can handle the complex TS data, which includes heavy data imbalances.

# 3 Related Work

As outlined in Chapter 2, ED in dairy cattle has been a top research subject in livestock management for the past few decades, transitioning from manual observation to the latest data-driven and automated systems. While the previous chapters explained the life cycle mechanisms and technologies behind the detection of estrus, this chapter presents a critical review of the earlier studies focusing ED, arranged chronologically that exhibits the development of technologies and the detection techniques. The section starts by introducing the traditional methods, continues with the advancement in automated sensor technologies, ML approaches, DL models, and latest innovations involving uncertainty estimations and development in analytical techniques through AI. Each section summarises the achievements, focuses on the comparative insights arising from the research, and eventually identifies the research gaps that the present thesis addresses.

## 3.1 Overview of Related Studies on Estrus

As one of the decisive factors which influences the dairy herd productivity, there has been a lot of research underway in the ED process. The earlier studies focused on getting detailed insights into the behavioural and physiological changes during estrus. It quantified the manifestations of estrus such as the mounting behaviour patterns, activity fluctuations, especially restlessness, vocalisations and mucous discharge. It was Senger [30] and Roelofs et al. [31] documented that fewer than half of all estrus events of high production cows have been captured using purely visual observation. The reliability of the outcomes of these methods has been reduced due to the short estrus duration and the increased silent estrus periods in modern dairy breeds. The need for automated methods to surpass the existing methods is clearly explained in the study done by Senger. ED remains one of the most important reproductive problems in the dairy industry, leading to substantial economic losses [30].

To overcome this subjectivity of decisions through visual observations, scientists investigated several ways, including monitoring hormone concentrations in milk and blood plasma. It helped to predict the time of ovulation in dairy cattle and hence help in the accurate insemination process of dairy cattle [32]. However, this approach tends to be very difficult in applicability because of the requirement of laboratory analysis, technicians, high cost and time consumption. Several studies utilising technologies such as palpation and transrectal sonography have been employed as part of reproduction management strategy with improved precision [33], later found to be impractical for large-scale operations.

These studies collectively conclude the limitations in manual and semi-manual methods, especially the limitations in deploying on a large scale for estrus monitoring and compromising the cost concerns. This outcome emphasises the need for integration of sensor technology into the reproductive management system.

## 3.2 ED using Sensor Technologies

With the rapid shift from traditional systems to automated methods, there have been different studies conducted utilising behavioural monitoring using sensor technologies. It is predominantly applied to track the behavioural changes undergone in cattle and to track the symptoms during an estrus period. Such studies are listed in the sections below.

### 3.2.1 Activity Based Sensor

The step-by-step progress towards automation can first be witnessed through tracking the activity. Activity monitors, pedometers and accelerometers have been widely used. Lovendahl and Chagunda [34] [35] made use of physical activity monitoring for ED in cows and demonstrated that the activity measured by pedometers increased approximately 2- to 3-fold during the estrus period, and had detection sensitivities approaching 90% [34] [35]. Eeman At-Taras and Spahr [36] also showed the efficiency of activity-tracking systems and highlighted the high ED efficiencies of 86.8% using an electronic heat mount detector and 87% using an electronic activity tag for the characterization of estrus, whereas visual observation achieved only 54.4%.

Later, the studies completed by Reith and Hoy [29] showcased the comparison between the devices mounted on the neck and leg, and concluded that the neck-mounted devices captured the behavioural changes earlier than pedometers. Although there have been several improvements in tracking the movements and proving the relation of it with estrus, the activity-only The monitoring system has the chance to produce false alarms because of changes in diet or social regrouping. Therefore, The system incorrectly flags the wrong stage as estrus [37]. This challenge opened a path to research more on behavioural indicators and to combine them with other measurements, such as physiological factors, which created a multimodal system.

### 3.2.2 Physiological and Multimodal Sensors

Maria Munkøe et al. [38] focused on evaluating rumination-time changes and observed the decrease in rumination during the estrus period. This study also evaluated the individual variability among the herds and observed the variations in different cows.

Similarly, the study done by Kathrin Henriksen et al. [39] explained the low correlation between rumination measurements completed through sensors and direct observations. It shows the good efficiency of sensors in overcoming direct-observation limitations [39]. The physiological factors similar to rumination have also been found useful — especially milk conductivity and electrical resistance — because both relate to hormonal fluctuations [40].

Rutten et al. [41] combined the activity, rumination and ear temperature into an integrated decision system and improved the sensitivity to 92% and specificity to 89%. To support real-time analytics, farms have been deployed with sensors like vitalimeter, which enables the continuous monitoring of physiological factors [24].

All these factors relied on the correlations with each other. The individual variability and the seasonal changes make these multimodal sensors challenging to adapt [27].

Although the studies show significant improvement when using multimodal sensor data, Rutten et al.'s study [41] of integrating multimodal sensors faced high false-positive alerts.

Consequently, researchers started focusing on fusing ML algorithms with multimodal sensor data, which is capable of studying the non-linear relations between the behavioural and physiological characteristics.

## 3.3 Machine-Learning Approaches for ED

ML approaches opened up the possibility of data-driven predictive modelling.

Wang et al. [42] demonstrated how ML techniques enhance ED by utilising location and acceleration data. They studied k-Nearest Neighbour and achieved around 73–90% accuracy. Back-Propagation Neural Network also showed accuracy up to 95% and Linear Discriminant Analysis up to 85%. The study concluded that ML along with wearable sensors produced better results than visual observation [42].

Shogo Higaki et al. [43] studied the feasibility of ML in behavioural sensor data collected by monitoring the ventral tail-base surface temperature. Behavioural data has been researched several times through different ML algorithms to find the best-performing one.

Malik Ergin et al. [44] have done similar research on behavioural data across seasons. They conducted the study on different algorithms, including Support Vector Machine, Random Forest and several other algorithms. It was found that Multivariate Adaptive Regression Splines performed better, with an accuracy of 0.95 and an AUC of 0.85 [44].

Although most of the studies compared across comparable algorithms, these analyses revealed some consistent drawbacks as well. It includes:

- The heavily imbalanced datasets, where estrus instances occurred less than 10% of total observations [42].
- The issue with generalisation capability, which made consistent weaknesses in the studies [44].

The imbalanced datasets even lead to a higher bias towards the majority class, and the differences in herd-to-herd behavioural characteristics also induced weaker results.

## 3.4 DL Approaches: Temporal Models

DL advances towards the heat detection process, especially by making use of temporal data provided by leveraging sensor benefits by monitoring specific behavioural and physiological characteristics. It includes the studies using models like RNN, especially LSTM networks, and hybrid models combining Convolutional Neural Network (CNN) architecture with temporal modelling structures. All these studies analyse patterns, help in detecting the estrus and predicting optimal insemination times. A significant contribution to the field was provided by Wang et al. [45]. They have studied and implemented DL based on estrus behaviour detection using CNN-based detection and YOLO detection pipelines. Their work used the important visual indication of estrus (standing heat), which is the mounting tendency showcased by cows, by capturing the posture keypoints of these cows [45]. This study highlighted the role of visual DL that can be integrated with sensor-based systems.

The studies collectively point out that by integrating the temporal information along with spatial details, DL achieves more precise estrus prediction.

### 3.4.1 RNNs

RNNs represents a general class of DL models designed for the processing of sequential information while retaining the hidden state information from previous time steps. LSTM networks form a type of architecture that enhances the basic RNN architecture with the concept of gated memory to improve the learning of long-term dependencies [46]. The development in estrus prediction brought by RNN and LSTM contributed by their peculiarity in modelling the temporal data, especially sequential or time-dependent raw sensor data patterns. The studies focusing on these models showcased significant progress in farm technologies. Chen et al. [47] showcased a direct implementation of a standard RNN for ED, whereas they have proposed an LSTM network which has been trained on 24-hour based behavioural data utilising the behavioural features. The model has achieved 0.95 AUC despite having a large class imbalance between estrus and non-estrus events. This study concludes that the memory-based modelling benefits the estrus detection than feed-forward networks. Beyond ED, researchers also focused on learning the reproductive transition, which also made a crucial impact in the studies for estrus prediction. The study done by Keceli et al. [48] used activity and behavioural data for calving predictions using RNNs. Since calving and estrus have been closely related, and indications are through the behavioural shifts, this study demonstrated that recurrent architectures are beneficial for reproductive forecasting.

Similarly, the study on the topic "Classification of cattle behaviour and detection of heat using sensor data" done by Druv Dakshinamoorthy et al. [49] focused on different models for effective detection of heat using sensor-provided data. Their study, using the LSTM model, correctly identified four heat days out of a total of 4 heat days in their dataset and identified 19 out of 20 non-heat days in their test set. This shows 96% of accuracy with only one false positive [49]. Although there have been several limitations in the study , such as the small evaluation set, the study provided an insight regarding the effective combination of LSTM with sensor data , which can be helpful in flagging estrus days. Thus, the studies stated above give several conclusions, such as:

- LSTM networks can be used strongly and precisely in the early class imbalance conditions over daily and multiple windows, even capable of achieving higher accuracy results only using accelerometer and such behavioural data from sensors [49] [47].
- RNNs are well-suited for the studies of reproductive forecasting, such as calving, indicating their implementation capability [48].

### 3.4.2 Why LSTM Networks?

Estrus behaviours include the following, which occur over time [50] [51] [52]:

- a gradual increase in activity,
- subtle decrease in rumination routine,

- changes in feeding,
- and multi-hour behavioural fluctuations.

These demonstrate either short cycles, especially variation from hour to hour or long cycles. In contrast to traditional ML, LSTMs are able to capture information about what has been observed through the memory gate [53] [49] [47]. Standard RNNs are known to have limitation in terms of vanishing gradients during backpropagation through time. However, this issue has been mitigated in LSTM networks through gated memory cells that allow information and gradients to pass through in order to ensure stable learning over long sequences [46] [54], so that they can:

- learn when changes in behaviour start,
- understand about long time patterns in the data,
- recognise historical context appropriately,
- and identify sequences showing similar profiles to estrus.

This is why LSTMs are suited to modelling the biological rhythm associated with the expression of estrus.

### 3.4.3 Deep into the LSTM

The LSTM, proposed by Hochreiter and Schmidhuber, was an attempt to rectify the issue affecting standard RNN networks, their weakness in handling long dependencies effectively in sequential data [54]. LSTM unit maintains two internal states: the cell state $c_t$ and the hidden state $h_t$ [53].

Both of them together make the network store and update data across time. The data flow can be managed by three mechanisms in the model [53].

**Forget Gate**

Decides which past data should be handled or removed:

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{3.1}$$

**Input Gate**

Determines which new information about data should be passed on:

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{3.2}$$

Along with the memory:

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{3.3}$$

**Output Gate**

Controls the data flow from the updated cell state to the next layer:

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{3.4}$$

**State Updates**
The cell state and hidden state :

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{3.5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3.6}$$

These formulas illustrate how an LSTM utilises the newest behavioural input and historical information carried from previous steps. The capability is critical for the detection of estrus, since more information is conveyed by the interaction of the behavioural variables over time rather than any single measurement.

## 3.5 Baseline Models

In this thesis, the Performance of the ED framework based on the LSTM model is contextualised by implementing two traditional machine learning models: LR and Linear SVM. The motivation behind choosing these models is based not on their superior performance in modelling sensor data, but rather because they are two well-established and theoretically complementary approaches to supervised learning. Both baselines share the same aggregate input feature representation as the LSTM model, thus facilitating a fair comparison that isolates the value-added component of temporal modelling.

### 3.5.1 LR

LR is a probabilistic linear classification model that estimates the posterior class probability. LR has been considered as a reference model in statistical modelling because of its ability to interpret, mathematical transparency and the robustness in weak signal and class imbalance settings [55].

Unlike more complex models, LR enables an easy link between input variables and predictions on the basis of the estimated coefficients [55]. This makes LR especially suitable as a reference model when understanding decision behaviour or probability calibration is of importance, as in the area of health-related or biological event predictions.

In the context of ED, the role of LR is that of a non-temporal and low-complexity model, which demonstrates how much discriminative information can be derived from aggregated behavioral features alone, without using temporal information.

### 3.5.2 Linear SVM

Linear SVMs are classifiers that aim to find a hyperplane maximizing the class separation margin. The success of Linear SVMs in high-dimensional spaces, their interpretable mathematics, and robust defences against overfitting make them a baseline learning model in classification

tasks. The literature has clearly emphasised the stability of Linear SVMs in moderate levels of class overlaps and additively dominant feature relations [56]. The Linear SVMs in the current work serve as a basic linear model to illustrate the separability capability of the feature space that could be derived in the absence of temporal structure and hierarchical feature learning processes.

**Reason for Using These Baselines:**

- Both LR and Linear SVMs represent two completely different yet popular linear classification techniques.
- LR provides a probabilistic solution with interpretative capabilities to show just how likely events of estrus occurrence can be determined from aggregated features by means of probability inference [55].
- The linear SVM focuses on the concept of margins for separation and reflects the separability of the feature space from a geometric viewpoint, not incorporating sequence information to provide a direct evaluation of the role of temporal representation learning provided by the LSTM model [56].

Both baseline models have the same input processing and representation mechanism as in the LSTM model. The role of these baseline models in the thesis is not to outperform sequence-based DL methods, but to offer clear, replicable, and tractably justified points of comparison to assess the value added by temporal representation and uncertainty quantification.

## 3.6 Summary

The summary of studies, along with the table 3.1, reveals a focused transition from individual-based visual observations to an automated data-driven and model-based heat detection. The researchers focused on establishing a close relation between the physiological and behavioural characteristics of estrus. Especially, sensor studies revealed the association of estrus signal with changes in activity, rumination, and temperature [34] [35] [36] [29] [37]. While the ML approaches showcased the possibility of exploiting non-linear patterns in this behaviour, despite the class imbalance and generalisation challenge [42] [43] [44], DL–based studies, especially those involving LSTM-based architectures, advanced the reproductive forecasting by learning behavioural long-range dependencies over the highly imbalanced conditions of the heterogeneity in the individual behaviours, the treatments of imbalanced data and the difficulty or challenge in early-level predictions. These challenges motivated the pathway to the evolution from traditional methods to automated predictions. Such gaps motivated us to focus this study on a pure LSTM-based architecture, supported along with the uncertainty estimations in the methods [47] [48] [49] [58] [57].

Table 3.1: Summary of ED Approaches, Data Sources, Methods, Findings, and Limitations

| Category | Studies | Data or Sensors Used | Methods Applied | Findings | Limitations Identified |
|---|---|---|---|---|---|
| Traditional & Manual Methods | Senger [30], Roelofs et al. [31] | Visual observation, behaviour signs | Manual inspection | less than 50% of estrus events detected, silent heats common | Labour-intensive, subjective, short estrus duration reduces accuracy |
| Hormonal and Veterinary Techniques | Van Eerdenburg et al. [32], Bagley et al. [33] | Plasma progesterone, milk hormones, palpation, sonography | Laboratory analysis | High physiological precision, improved ovulation timing | Expensive, slow, impractical for herd-scale daily monitoring |
| Activity-Based Sensors | Løvendahl & Chagunda [35] [34], At-Taras & Spahr [36], Reith & Hoy [29] | Pedometer, accelerometer | Activity indexing with threshold-based alerts | 2-3 times activity increase during estrus, sensitivities up to 90% | High false positives (diet change and regrouping), behaviour varies between cows |
| Physiological & Multimodal Sensors | Henriksen et al. [39], milk conductivity studies [41], Rutten et al. [17] | Rumination sensors, temperature, milk conductivity, ear sensors | Multimodal fusion, heuristic rules | Rumination decreases during estrus, multimodal systems improve sensitivity (92%) and specificity (89%) | High cow-to-cow variability, environmental influences, false-positive alerts |
| Classical ML Approaches | Wang et al. [42], Higaki et al. [43], Ergin et al. [44] | Accelerometers, tail-base sensors, seasonal datasets | SVM, Random Forest (RF) | Accuracy up to 95%, ML captures nonlinear relations | No temporal modelling,MLsevere class imbalance, weak generalisation across cow |
| LSTM and RNN-Based Models | LSTM Estrus Model (Chen et al. [47]), Calving RNNStudy | Behaviour sequences (24h), multi-day physiological & activity data | LSTM, stacked RNN | AUC up to 0.89, strong ability to capture long-term behavioural deviations | Imbalanced data, limited evaluation across multiple farms |
| DL Models Beyond Pure RNNs | NB-IoT LSTM Hybrid [57], ML-Hybrid Baseline [58] | IoT behavioural streams, multivariate activity, ML-classified activity | CNN-LSTM, multistage hybrid pipelines | Sensitivity up to 94%, improved feature extraction | Computationally complex, not adopted in this thesis, limited generalisation research |

# 4 Methodology

The methodology has been designed in collaboration with the company Farmtec a.s. The company has proposed the LSTM network as a solution for early ED because of its suitability to the TS data. This data will be collected and preprocessed to ensure its quality and consistency. The chapter focuses on building a complete pipeline by compounding the biological and technical information collected from the previous chapters.

The methodology chapter sticks to a structured workflow. Beginning with the farm setup and data collection procedure being explained, the chapter outlines the pre-processing and feature engineering principles followed in the implementation. The methodology is prominently devoted to the design and implementation of the LSTM model with different prediction horizons. Along with the descriptions of the study setup and the different modelling strategies, the chapter concludes with the estimation of predictive uncertainty, which is relevant at the time point of reproductive decision-making.

## 4.1 Farm Setup and Data Collection

The dataset used in the study was provided by Farmtec a.s., the industrial partner of the study. They operate and generate the modernised livestock technologies for dairy farms. They have integrated technologies, especially sensor technology, including vitalimeter, to enable the hourly based monitoring of each cow in the farm. The data collected are visualised and accessible through a farm management application. This helps the farmers to analyse the trends of the behavioural data. The data collected through the smart neck collar (vitalimeter) is stored on Farmtec a.s. internal company server. Users access and operate the system via the online application vitalimeter.com, which serves as the primary interface for viewing dashboards and managing the data. The specific farm operational and business details remain confidential. The overall data-collection process and the automated monitoring of each cow showcase a modern, innovative dairy management system.

### 4.1.1 Behavioural Tracking and Indicators

Farmtec a.s. has deployed this real-time monitoring system for cows in the farm on an hourly basis, which helps to track the behavioural and physiological measurements continuously. The ordered hourly representations, along with the numerical categorization of individual herds separately, enable a seperation between each cow's data. It helps in not having mix-up of data between different herds. This system together enables the shifts in the values and evolving patterns to be captured over time. This creates a profile for each cow tracking the behavioural shifts with the changes in intensity and stability of transitions in the cow's physiological state, indicating the chance of estrus.

The dataset consists of TS measurements of how each cow moves, and maintains the feeding, as well as the intensity of the rumination done by the cows. This characteristic has been observed over the days to find the characteristics exhibiting a reproductive phase, the estrus in the dataset is expert-labelled. Besides this sensor-captured information, general cow-level background information is taken. The reproductive status and age category information provide contextual knowledge about the herd for understanding the individual behavioural variations.

Among the behavioural characteristics mentioned, three categories play a central role in estrus due to their consistent physiological involvement in the reproductive phase. Firstly, the herds increase their physical activity up to four times, especially the number of steps taken per hour [50]. The activity-related measurements show a significant shift or sudden spikes as the herds become more restless and socially stimulated while having estrus.

Secondly, rumination accompanying food intake in herds also tends to show interruptions and reductions. This showcases the hormonal adjustments that give insights into estrus [51]. Thirdly, the change in behavioural priorities can be observed through the feeding-related characteristics in cows. It also exhibits the sudden declines and irregularities in the feed intake individually [52]. Analysing this behavioural and physiological characteristic gives partial insights. However, those characteristics observed together will provide meaningful signatures for the biological processes underlying the estrus. By evolving meaningful sequences in the pattern of these behavioural shifts, the indicators point out the estrus onset.

Figure 4.1 shows the behavioural changes during estrus period , which points out the estrus pattern shifts in the indicators. This time-dependent behavioural shifts are integrated with the sequential modelling approach developed in this thesis, to enable the LSTM architecture to learn the temporal characteristics and the information pattern between activity, rumination, feeding and estrus prediction approach.

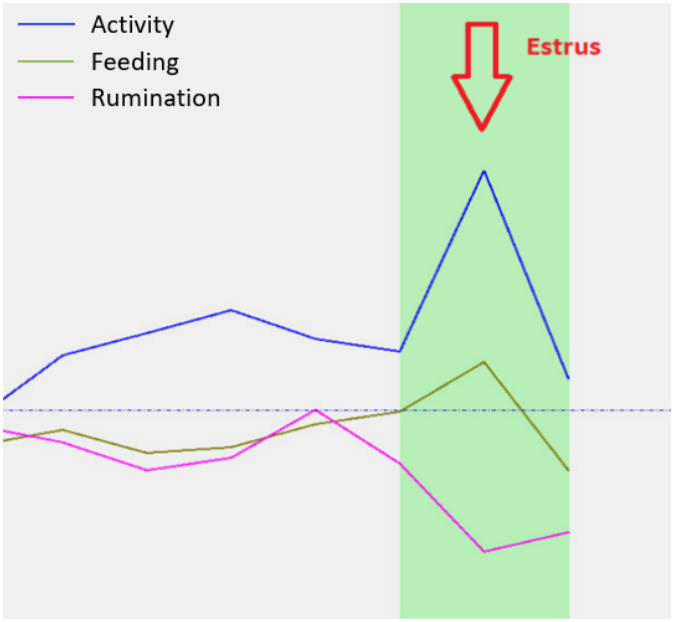


Figure 4.1: Behavioural shifts during estrus (provided by Farmtec a.s.)

## 4.2 Development Environment, Tools, and Frameworks

The development approach of this thesis is drafted to efficiently support the easy and optimal handling of data, the Processing of sequential data, and the execution of uncertainty-based experimentation. In this case, the entire process of testing and analysis in the models and methodologies utilised have been implemented in Python and executed in Jupyter Notebook hosted on the Visual Studio platform that supports exploratory series analysis steps for transformations and Evaluation procedures.

## 4.3 Data Sources and Format

The type of dataset that is used in this thesis is Comma Separated Values (CSV) files generated from Farmtec a.s.'s behavioural monitoring system. CSV files contain hourly behavioural recordings for 296 cows, along with the factors such as activity, rumination, feeding and reproduction status markers, along with generic details such as age category. The Entire Processing, including data loading and data pre-processing, was carried out on the CSV files via Python's pandas library. It supports strong-index-based merging and grouping, along with cleaning, with suitability to sequential models [59].

## 4.4 Tools and Libraries

A full suite of Python libraries facilitated the entire modelling workflow, starting from preprocessing and sequence creation to the construction of deep models, as well as calculations for estimates of uncertainty. TensorFlow and Keras were used for the LSTM architecture, associated layer definition, dropout, model compilation, train loops, and predict functions. Such libraries offer acceleration via GPUs, along with ample flexibility that suits time-based deep models.

Python libraries such as NumPy and pandas were used for numerical computations and manipulation of the data. Scikit-learn libraries were providing traditional ML solutions, train-test-splitting solutions, handling imbalanced classes, hyperparameter tuning, tools and metrics such as precision, recall, F1, ROC-AUC metrics.

Throughout this project, Matplotlib and Seaborn libraries were used for visualization purposes, such as plotting TS series data, sequence distributions, windows of predictions and histograms of uncertainty calculated via the MC dropout.

## 4.5 Workflow

The methodological workflow of this thesis consists of a sequential approach of stages that will allow the raw behavioural observations to be modelled as hourly predictions on the estrus and the measures of quantified uncertainty. This methodological procedure includes the preparation of data, including the pre-processing, high-level feature engineering, sequentialization of behavioural observation and LSTM modelling.

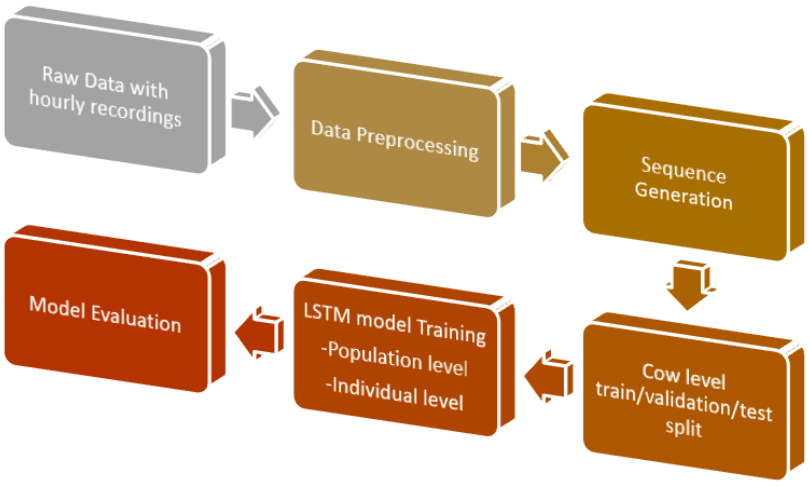


Figure 4.2: Work flow

A schematic representation of this workflow is shown in Figure 4.2, which visualises the key elements within this analytical workflow. This workflow begins with importing raw hourly behavioural recordings derived from the monitoring system within the farm. Following preliminary processing, including data cleaning and formatting, the data undergoes sequence preprocessing, such as normalization and the establishment of higher-order temporal features. These phases of processing remain intentionally carried out at a generic level according to the requirements imposed by the company's confidentiality. This pre-processed form of the dataset will then be further sorted chronologically in terms of individual cows and prepared for sequential modelling by creating fixed-length windows of input.

After organizing the behavioural histories into sequences, the task of ED can be defined as a sequence classification. Each sequence represents a fixed prediction horizon, either the current estrus condition, or the estrus condition of the cow in the future 3 , 6 or 12 hours. These sequences are associated with labels generated based on the tracked estrus condition. After this, a cow-level split is used, which divides the dataset into train, validation and test sets. This ensures that the behavioural histories of cows within the test set will not be seen in any form through the entire model training process.

Model development has been conducted through two different evaluation approaches,

1. A population-level LSTM model that focuses on shared cow behavioural dynamics.
2. An individual-level LSTM evaluation approach, model is assessed at individual cow level.

Although the same structural design is used in both cases, their focus differ in how performance is evaluated across cows. In population level approach, the train, validation, and test split is done in a way that the entire set of sequences of cows in the test split is completely unseen during training and validation, which allows a non-biased estimation of generalization performance. In contrast, for individual-level assessment, the data from all cows is taken for

each split while maintaining the temporal ordering within each cow, allowing performance assessment of the same trained model for each cow separately. Population level evaluation highlights overall system performance on unseen cows, while individual level evaluation reflects the model performance variability across cows under similar training conditions.

## 4.6 Data Preprocessing

The pre-processing phase helps in preparing the raw behavioural data from the real-world farms into a clean and consistent form suitable for sequence-based modelling. Taking into consideration that the dataset was collected from real-world farms, some natural occurrences such as missing information, noisy behaviours and real-world hourly variations are expected to be visible in the data provided. However, a series of high-level pre-processing phases was carried out to keep intact the integrity of the information and meet the objective that all modelling phases will be supported with reliable inputs. These phases will comply with industry best practices for TS preparation, without revealing any proprietary information on their handling of the data, which was provided by Farmtec a.s.

### 4.6.1 Data Cleaning and Structural Preparation

The first stage of pre-processing was concerned with ensuring that the record of behaviour for each cow formed a continuous, structured TS. The Some general principles of cleaning were carried out, such as ;

- Treatment of missing data with conventional TS methodology to ensure uninterrupted sequence generation.
- Removal or smoothing of irregular entries, which means the filling gap between the data that may occur due to the temporary loss of sensor signals, communications and environmental disturbances.
- Maintaining chronological order by ensuring the timestamp orders of recorded date so that the sequence of behaviours for each cow can be computed without any temporal inconsistencies.
- Filtering entries that have partially or largely incomplete behavioural profiles, as their lengths will not meet the minimum requirements posed by the LSTM structure.

All this was done with expertise with reference to the conventional scientific methodology in preparing TS for behavioural studies and made sure not to reveal the internal workings of the firm.

### 4.6.2 Normalisation and scaling

To aid with stable optimization, variables representing behaviours were transformed through the use of standardised normalization. Normalization promotes equal weighting of variables

with significant differences in magnitude (for instance, activity monitored values and rumination time) as contributions towards the model's representations.

Although the specific transformation functions will not be shown due to confidentiality, the process did follows principles, such as:

- Scaling behavioural features to identical ranges,
- Lowering the dominance of features with large numerical scales,
- Maintaining temporal structure without modifying the form of behaviours, and to attain uniformity amongst the cows. These schemes for normalization are consistent with conventional sequential models and allow for effective optimization through gradients in deep neural networks.

### 4.6.3 Feature Engineering

In feature engineering, further behavioural and temporal information was derived from the basic raw measurements provided. Although the nature of feature engineering is proprietary in accordance with Farmtec a.s., this includes the types of transformation that will be mentioned further below. These consist of:

- Short-term behavioural details that capture behavioural shifts over recent hours.
- Long-term behavioural summaries implying the gradual development of a trend or pattern.
- Relative behavioural changes that refer to deviations of present behaviour from the cow's usual behaviour pattern.
- Lagged behavioural features, created by including previous hour values to provide temporal context for non sequential models.
- Temporal contextual markers, such as those that show day or periodic cycles.
- Aggregated multi-scale indicators, that merge the short term with the long term for better representation.

Such designed categories allow the model to train on changes in behaviour over different time scales, which is essential for heat detection. However, the specific transformation rules, variables, and numeric ranges remain unexposed to protect the secrecy of the intellectual property of the company in feature processing.

### 4.6.4 Sequence-Based Learning

After cleaning, normalization and transformation of the behavioural data into structured representations, the dataset was ready for sequence-based modelling. The procedures include specific criteria such as

- Organizing behavioural records for individual cow.
- Providing fixed-length windows of time.
- Aligning sequences with prediction horizons, and generating labels for different estrus labels at different future time intervals.

Such top-level steps guarantee compatibility with the LSTM architecture, as LSTM models expect their input tensors to be ordered in time.

Details of the sliding window, internal thresholds, and proprietary segmentation schemes remain unspoken as they relate to elements within the protected pipeline used by the partner company. However, will focus on principles supported within academics with regard to the transformation from raw behavioural observation into a structured form.

## 4.7 Problem Formulation and Sequence Generation

The problem of ED in dairy cows can be termed as a sequence-classification problem, which has the goal of classifying the estrus state of an individual cow based on the sequence of behavioural patterns. Since the onset of estrus behaviour occurs not incidentally, but as a pattern over time, the modelling strategy needs to address the behavioural patterns as they develop over the period of an hour. The basic concept is that for every individual cow, the sequence of her behavioural patterns needs to be translated into a sequence of similar length. Every sequence contains the behavioural patterns recorded for consecutive periods of an hour. This forms the inputs for the model and are labelled according to the individual cow's estrus state.

### 4.7.1 Prediction Windows and Target Variable

Detection of estrus not only involves finding the present status, but also needs the prediction of the impending onset of estrus. The predictions in advance is crucial for the accurate and timely inseminations to ensure the continuity of the reprodutive phase.

In such predictions, the following four prediction horizons are considered in the study.

- current estrus cycle phase (0-hour window)
- 3 hour prediction
- 6-hour prediction
- 12-hour prediction

These horizons are specified relative to the last step of the input sequence.

Estrus prediction is framed as a binary classification task:

$$y_t^{(w)} \in \{0, 1\}, \tag{4.1}$$

where 1 denotes estrus and 0 denotes non-estrus at window $w$.

For a window ending at time $t$, the prediction target for horizon $w$ is defined as:

$$y_t^{(w)} = \mathrm{E}(t + w), \tag{4.2}$$

where $w \in \{0, 3, 6, 12\}$ and $E$ is the estrus status at $t + w$ time .

This formulation makes it possible for the model to learn not only the behavioural signature of estrus but the behavioural precautions that begin to emerge well before the onset of the estrus period. Having multiple horizons makes it possible for direct comparisons to analyse the degree to which the behavioural patterns influence the timing of estrus and the degree to which estrus can be predicted.

The model is optimised through loss functions, which are suitable for handling imbalanced datasets. This way, the model learns to differentiate estrus from non-estrus over different prediction windows. The specific methods and loss function to handle an imbalanced dataset are going to be tailored in the study, which can be explained later in the upcoming sections.

### 4.7.2 Sequence Construction

When formulating label-aligned sequences, it is important to respect the time structure. For each cow:

- Sort the behavioral TS in chronological order.
- Generate sliding windows of length 'L'.
- For every window that ends at time 't', assign the label based on the chosen 'w'.

Windows near the dataset edges that cannot be adequately labelled are discarded. This guarantees that every training input has a clear and specific target. The sequences reflect the most recent behavioural trajectory up to a potential estrus event, and the labels indicate the time at which estrus occurs in such a way to determine whether it occurs at or after the sequence boundary.

To avoid information leakage across data splits and keep things biologically realistic, the sequence generation preserves the natural order of events. It ensures that the data across different cows will never be crossed. This approach allows the model to capture not just the typical behaviour during estrus, but also the behavioural changes that appear hours before the heat begins. By looking across several time windows, we can directly compare how patterns unfold over time and how early an estrus can be reliably predicted.

### 4.7.3 Dataset Partitioning for Sequential Modelling

For population-level modelling, the data is split into training, validation, and test sets with a split based on individual cows. This ensures that a particular cow appears only in one split. This design guarantees that there is no data leakage, which would create a chance for the model to memorise cow-specific characteristics instead of learning general estrus-related dynamics.

A representative split used in this study consists of a ratio :

1. Training set: approximately 70% of cows.

2. Validation Set: approximately 10% of cows, the held-out cows for hyperparameter adjustment.

3. Test set: approximately 20% of cows used for evaluation.

By this division, the Cow level split is enabled, and this reflects the model's ability to generalise on profiles that have never been observed before, it satisfies the crucial requirement in practical ED models. This avoids overlapping of individual behavioural characteristics and guarantees that the test performance measures generalisable behaviour rather than memorization. Cow-level split is important in behavioural modelling because cows have unique patterns in feeding patterns, activity rhythms and rumination tendencies that should not mix between training and testing. For individual-level evaluation approach, data split has been done chronologically within each cow into training, validation and test sets approximately 70%, 10% and 20% respectively.

## 4.8 Model Design and Implementation

The behavioural symptoms unfolding before the estrus cycle take several hours and often indicate a progressive transition rather than a sudden occurrence. These dependencies cannot be captured by models that treat the observations independently, without considering the historical relations and combinational spikes of features. Hence, the modelling approach should capture the temporal dependencies, and the study is based on LSTM networks. It is a type of RNNs tailored to learn such dependencies by retaining information over long time periods. LSTMs are useful tools to recognise estrus because they are able to accumulate behavioural paths [49], such as increasing peaks of activities and reducing rumination into a representation. The adopted approach is expected to predict how such behaviours tend towards estrus.

The design of the models in this work is built on two differing approaches. Firstly, population-level modelling that is based on learning behavioural patterns shared across the entire herd of cows. It is assessed by training model on a subset of cows and evaluating it on seperate set of completely unseen cows. Secondly, the individual-level evaluation method, which the data partitioning has been done chronologically on each cow and trained on pooled training segments across all cows. The model performance is assessed based on the pooled training segments, where the model makes predictions about the future outcomes of estrus in individual cows based solely on their past measurements while preserving temporal nature.

### 4.8.1 Architecture

The LSTM network designed should be capable of capturing temporal patterns related to estrus while functioning effectively even when faced with a high level of class imbalance or scaled data. All these considerations were based on domain expertise and experimentation.

**Input Representation :**

The input samples are modeled as fixed-length temporal sequences that have:

- Sequence length, $L = 24$ (24 hourly observations)
- Features, $F = 14$ behavioural and contextual variables

The choice of the window was based on the biological cycle of estrus, whose typical dynamics are measured within the Daily Time Window of 24 hours. Using the shorter time windows would lose the context, and longer windows would result in redundancy.

**Recurrent Layers (Stacked LSTM) :**

The architecture is composed of two layers of LSTM:

- First LSTM layer with $U = 64$ units
- Second LSTM layer with $U/2 = 32$ units

This setup is capable of doing hierarchical temporal learning, with the first layer learning the fluctuations of activity in the short term and the second layer learning the patterns in the higher level. The number of units was set using the hyperparameter tuning results, with larger configurations not being beneficial and at risk of overfitting.

**Regularization (Dropout) :**

The dropout rate of 0.4 is used after each LSTM and dense layers. The rate of 0.4 is determined after analysing the results through different executions and reflective of the fact that a strong regularization is required in this case due to:

- Less estrus events.
- High temporal correlation among sequences.

Moreover, the incorporation of dropout allows for the estimation of uncertainty in the later stages through the use of MC Dropout, explained in the upcoming sections.

**Fully Connected layer and Output Layer :**

A fully connected dense layer with 4 units and ReLU activation function is employed to extract compact non-linear representations from the temporal features. A small size for the hidden layer was chosen to deliberately keep the model capacity low. The output layer has one sigmoid-activated neuron, which estimates the probability of occurrence of estrus. The resulting network has 32,777 trainable parameters.

**Model Selection Rationale :**

The final architecture and corresponding hyperparameters were established using controlled experiments, which were performed using a predetermined data split. Comparisons and selections primarily driven by metrics such as the Area Under the Receiver Operating Characteristic Curve (AUC), because of its threshold-independent nature and suitability for imbalanced datasets. The support metrics such as precision, recall, and the corresponding F1 score were considered to improve explanation and were not used for optimization.

### 4.8.2 Population Level vs Individual Level Modelling

In order to evaluate the different strategies for the detection of estrus, this thesis describes the analysis of two methods: population level modelling, as well as individual modelling. While population models attempt to establish a general pattern of behaviour for the entire group of cows, individual models focus on cow-specific temporal patterns by evaluating model performance within each cow. Table 4.1 showcases the conceptual comparison between two modelling approaches.

Table 4.1: Comparison Between Population-Level and Individual-Level Modelling

| **Population Level Modelling** | **Individual Level Modelling** |
|---|---|
| Evaluated on unseen cows | Evaluated within seen cows |
| Cow wise data split | Temporal split within each cow |
| Evaluate generalisation to new cow | Evaluate temporal consistency within cows |
| Estimation of global performance | Reflects performance variability across cows |

## 4.9 Prediction Windows and Label Assignment

The thesis is not limited to the prediction of current estrus status, but also includes a number of potential times in the future. This showcases the real farm conditions, such as the farmer's requirement of getting advance information about estrus to plan effective insemination at the right time and set up other management processes, such as organizing the labourers. Hence, model assessments focus upon the current state and a number of hours in the future. The hours are measured relative to the end of each input pattern and will include the current state (0-hour prediction), and will include other hours in the future, such as hours 3, 6 and 12 into the future. The process will utilise the same model to answer a number of related questions, such as the current and future estrus status.

The temporal offset of labels has two primary purposes regarding the dataset. Firstly, it ensures that each sequence is associated with a well -defined and meaningful target. Sliding window focusing on the latest activity is consistently matched to a particular decision point in the future. Secondly, the sequences close to the edges of the temporal scope of the dataset, for which there is no appropriate target in the future and cannot be determined adequately for a particular window, will be removed systematically to ensure that the training and test samples clearly refer to a well-defined estrus and non-estrus phase at the corresponding decision point in the future.

The fact that the model has to work with more than one prediction window affects the level of difficulty in the prediction task. For the 0-hour window, the variations in estrus regarding activity, rumination, and feeding are expected to be more significant , and the model has access to the behaviours that are closer to the target occasion to a great extent. For the longer prediction horizons (6 hours and 12 hours), The behavioural cues are less prominent, and more variations in the behaviours of the cows are expected to be recorded. This makes the classification difficult, and the chances of misclassification of the estrus phase with the normal phase will be higher. analysing these windows side by side will overcome this risk in classification to reliably predict the estrus period.

In terms of modelling, the multi-window approach is applied to the LSTM model structure with no changes to the architecture. The same processes involved in the generation of sequences and feature engineering are followed in this model, with the only difference being lying in the manner in which the labels are assigned to each sequence of the model output.

The advantage of the multi-window model in terms of functionality is that it allows for an effective comparison of the model performance in terms of the window, along with the employment of the same model structure and representation. The model allows for the calculation of a uniform performance measure across the horizons in terms of metrics such as the value of AUC and F1-score. In order to measure the predictivity loss at any given window, depending on its distance to the estrus event in the timeline.

Taken together, it provides a basis for a comparative assessment involving short-term and medium-term predictions in preparation for the following analyses of model performance and its corresponding uncertainty regarding a variety of forecasting horizons.

### 4.9.1 Hyperparameter Tuning and Final configuration

The final LSTM architecture setup was obtained by systematic hyperparameter tuning. The variables tested were:

- length of sequence (timesteps)
- number of LSTM units
- dropout rate
- learning rate
- Batch Size

- number of training epochs

Analytically, the tuning procedure involved following conventional DL optimization techniques while incorporating the necessary constraints associated with modelling behavioural data. The obtained set of tuning attributes is standard across various prediction horizons.

## 4.10 Training Strategy and Handling Imbalance

The LSTM model training for the detection of the estrus needs particular attention to both the inherent properties of the data, especially the biological characteristics, and the underlying statistical properties posed by its class distributions.

This class imbalance problem, along with its temporal characteristics, points out the importance of approaches for efficient training that should be based on the temporal semantics of data, not biased towards predicting the majority class. The model training strategy used in this thesis focused on methodological principles to handle the class imbalances and hence stabilise the performance.

### 4.10.1 Class Imbalance in ED

In estrus behavioural tracking systems, estrus events are only a minor part of the daily and weekly behavioural pattern of a cow. considering the total number of non-estrus hours, this leads to a class-imbalanced data set where only a small proportion of data belongs to the positive class. As most hours belong to non-estrus conditions, a ML model would more likely forecast a solution where predictions consist entirely of the negative class or majority class. This might be a critical issue because a ML model could accurately classify most of the samples but might entirely miss estrus events.

The estrus cycle imbalance issue is complicated further by:

- Occurence of several estrus hours in time around estrus windows.
- Cow-specific behavioural trends with varying intensity and pace.

This necessitates a training approach that can focus the model's responsiveness to infrequent patterns while maintaining model stability. Table 4.2 summarises the class distribution along with the existing class imbalance in the dataset.

Table 4.2: Class distribution in the ED dataset

| Class | Count | Share (%) |
|---|---|---|
| Non-estrus (negative) | 1,945,089 | 99.19 |
| Estrus (positive) | 15,951 | 0.81 |
| Total | 1,961,040 | 100.00 |

### 4.10.2 Class Weight Strategy

The Sparse density minority class leads to a significantly skewed class Distribution of estrus labels. While training a ML model with a class-imbalanced data Distribution, ML algorithms bias towards learning more about the majority class. This often leads to a commendable level of overall accuracy but fails when dealing with a minority class. This observation has been acknowledged in ML literature. Here, conventional ML models are observed to overvalue the majority class and misclassify minority cases more frequently [60].

However, to overcome this class imbalance problem, this thesis adopts a class weighting approach when training the LSTM. This means that class weights are assigned to samples based on class prevalence. As explained by Bakır Arar and Elhan in 2023 [60], Class weighting belongs to one of the most successful methods for a rare event classification task because it. "gives more importance to misclassified samples of the minority class and deems the majority class errors less important".

Class weights are directly calculated based on the label distribution in the training data. Let us assume that we have $n_0$ number of samples belonging to class 0, and '$n_1$' number of samples for class 1. A typical weighting scheme introduced in literature for this particular classification problem [60] is :

$$w_k = \frac{1}{n_k}, \tag{4.3}$$

Here, $w_k$ represents the weight and $n_k$ represents the number of samples given to class $k$.

A more robust substitute for such a weighting function would be the inverse square-root weighting , which can be represented as [60]:

$$w_k = \frac{1}{\sqrt{n_k}}. \tag{4.4}$$

Which moderates the aggressiveness of weights while still increasing minority class sensitivity [60].

During the training process using TensorFlow, class imbalance is handled by incorporating class weights into the training process. This is achieved through the training process where loss is computed, such that loss for estrus-positive sequences is given more weight compared to loss for non-estrus sequences. Therefore, the loss arising from the misclassification of the minority class is given more weightage in the total loss computations. This helps in encouraging the learning of discriminatory features for the estrus event that is less frequent. In regard to ED classification problems, the importance of boosting the minority class cannot be overstated. Omitted estrus cases translate to unsuccessful insemination chances and lost revenue. As such, the correct classification of the minority class is more important than the accurate classification of the majority class. Weighted class classification would be a vital tool for focusing model classification power towards estrus sequences.

### 4.10.3 Focal Loss

Just as class weights focuses on imbalance issue, there is another common issue underlying imbalanced labels, the easy negative labels, which is defined as the samples belongs to the

majority class that the model can classify with high confidence. This is because most sequences belonging to a non-estrus class are easy to classify for a model. Therefore, this dominance leads to a focus of the optimization process towards easy sequences. As a result, fine behavioural characteristics of pre-estrus sequences may be overlooked.

To overcome this problem, this study integrated with Focal Loss, a cross-entropy loss function modified to downplay the contribution of well-classified samples biases learning towards challenging and informative samples. Focal Loss was introduced by Lin et al. [61] for dense object detection tasks. The presence of a vast class imbalance between the foreground and background classes prevented convergence. Nevertheless, their experiments clearly indicated that learning with complex samples has a vast potential to increase the minority class precision without adding any computational complexity.

The focal loss for binary classification is :

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t), \tag{4.5}$$

where $p_t$ is the predicted probability assigned to the actual class, $\alpha_t$ is the balancing factor between minority and majority classes, $\gamma$ is the focus parameter that manages the degree of weighting of easy samples.

Focal Loss imposes a trade-off between the minority class and majority class, and ɣ represents the focus parameter, which controls the degree to which easy samples are down-weighted. A larger $\gamma$ means that the model tends to focus more on samples with low frequency instead of being misled by samples with high frequency. This tends to be more effective in ED because pre-estrus cases only displays slight differences in their behaviour that can be overshadowed by the abundance of unimportant non-estrous cases.

Later developments in focal loss have shown that their designs are universally applicable to various types of imbalanced datasets. Batch Balanced Focal Loss was introduced by Singh et al. [62]. This model integrated a focus mechanism with balanced mini-batches achieved better binary and multi-class medical image classification. The application of Focal Loss in this thesis has backing based on both theoretical and practical foundations. This occurs because estrus behaves as a physiological phenomenon that occurs with low frequency. Focal Loss can be employed to ensure that LSTM focuses more on such loose behavioural variables. Focal Loss can be used to ensure that easy estrus samples are disregarded during training. The parameter $\gamma$ used in this thesis has been considered within values shown to be effective within previous experiments. The model is employed with both class weight and focal loss together to address the above mentioned issues.

### 4.10.4 Optimisation

Model optimization is done using the Adam optimiser. It is employed with a learning rate of $1 \times 10^{-4}$, default parameters and with a clip norm of 1. This has adaptive gradient calculation capability that works well for LSTM neural networks. Additionally, for model stability during training, dropout regularization is used in the LSTM layers. This is for the regulation during training, which causes overfitting due to the cow-specific noise.

The training occurs in a mini-batch with a constant number of samples and multiple epochs until convergence. Also, validation loss and appropriate evaluation metrics (like F1-score and AUC values) are employed for stopping training and hyper-parameter tuning.

## 4.11 Uncertainty Estimation

Uncertainty estimation is an important aspect in estrus prediction, and it becomes even more challenging because of the decision boundary. In discriminating between estrus and non-estrus actions, it is uncertain which is characterised by noise and significant variations across the cow population. Conventional DL-based classifiers make point predictions in a deterministic forward pass, providing no insight into the confidence level of the model in the prediction made. Reproductive management in a cow herd requires confidence in the model since a false positive leads to unnecessary inseminations, while a false negative results in lost breeding possibilities. The uncertainties can be the inherent noise in measurements or the noise in model parameters due to the lack of training data. Estrus behavioural observations contain both natural factors, such as sensor noise contribution, and the class imbalance of estrus and non-estrus hours. For this purpose, without any modifications to the basic structure of the LSTM network, the current study proposes the application of the MC Dropout technique. It is a dropout methodology put forward by Gal and Ghahramani in their work (2016) [63]. The process of dropout is viewed from a Bayesian approximation perspective, which makes it an effective way to calculate the model's uncertainty based on the dropout layers used in model training. In practical perspective, the predictions with high uncertainities have been considered as low confidence predictions and they indicate a high chance of misclassification. In such cases, the insemination should not to be performed without verifying any additional observations related to estrus to avoid unnecessary expenses.

### 4.11.1 MC Dropout

The MC Dropout model allows for an efficient calculation of the Bayesian inference in DL methods. Unlike regular deep networks, where the dropout is turned off during testing, the MC Dropout model keeps the dropout active and performs a number of forward passes. Each pass, make sure to sample a different network by dropping random units and thus samples from the approximate Bayesian posterior distribution of the network weights [63].

Gal and Ghahramani (2016) proved that the application of dropout layers in front of each weight layer has an identical form to the application of variational inference in deep Gaussian processes. It makes the application of dropout a form of regularization. The fundamental insight behind the application of dropout is the fact that it implicitly samples from a varying distribution [63]. for an input $x^*$, the predictive distribution is :

$$p(y^* \mid x^*) \approx \frac{1}{T} \sum_{t=1}^{T} f(x^*; W_t), \tag{4.6}$$

Where each $W_t$ is a set of weights obtained from a stochastic forward pass.

**Predictive Mean**

$$\hat{\mu}(x^*) = \frac{1}{T}\sum_{t=1}^{T}\hat{y}^{(t)}, \tag{4.7}$$

**Predictive Variance**

$$\hat{\sigma}^2(x^*) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{y}^{(t)}\right)^2 - \left(\hat{\mu}(x^*)\right)^2. \tag{4.8}$$

Within the current thesis, the application of MC Dropout to the trained LSTM model involves making $N$ stochastic forward passes per hour in the test set. The variance of the predictions made will be a measure of how confident the model is:

- Low variance represents high confidence,
- High variance implies the presence of uncertainty, often linked to unclear behavioural cues and limited training examples relating to instances of estrus.

As the values for the standard deviations from MC dropout are continuous values that do not necessarily take a normal distribution, a non-parametric Mann-Whitney U test is used for determining whether the distribution of uncertainty for incorrect decisions was stochastically higher than that for correct decisions [64]. To analyse the monotonic relationship between the predictive uncertainty and classification results, Spearman's rank correlation coefficient and the probability value(p-value) for statistical significance were used. The correlation coefficient varies in between $-1$ to $1$ where $0$ indicates no correlation and p-value less than $0.05$ considered to be statistically significant. It should be noted that the choice of the Spearman's rank correlation coefficient as the method of choice for analysis is based on the fact that it does not assume linear and normal distributions of data since the ED is highly imbalanced [65] [66]. This uncertainty information is then used in the evaluation chapter to describe the prediction reliability, the risky hours of prediction, and additional information based on the deterministic classification metrics.

## 4.12 Individual Cow Metrics for Individual Level Evaluation

The evaluation of estrus prediction models based on individual-cow metrics is crucial since the pattern of estrus behaviour tends to vary significantly across individual animals, and population-level metrics might hide the variations in the individual cows. Even if the population-level metrics such as the F1-score and AUC provides an average performance level of the model, but it does not offer insight into the model's performance level for each individual animal. Because it tends to perform well for some and poorly for others. The individual cow metrics are not reported for population-level modelling, as testing is performed on completely unseen cows and focused on aggregated performance rather than variability across cows.

Estrus expression is affected by a number of cow-specific variables such as age group, lactation, parity, temperament, feeding pattern, and variation in rumination. Since the LSTM

model is trained under a temporal partitioning strategy, it is important to check if the predictions made by the model varies across cows. The per-cow assessment will take care of this issue by understanding the performance variability across cows, which allows identification of uneven perfomance rather than focusing on generalisability.

The assessment framework provides per-cow results in terms of F1-score, precision, recall, and AUC values per cow. The values are measured by considering only the test instances in the dataset corresponding to a particular cow and ignoring any cross-cow results. The results will provide a detailed insight into the performance of the model in identifying estrus in each cow consistently.

Moreover, the per-cow assessment helps to identify the possible presence of bias due to the differences in the number of sequence observations. The presence of fewer observations concerning estrus occurrences leads to fewer instances for the positive samples, and this might significantly impact the performance relative to the other cows with balanced behavioural data.

## 4.13 Baseline Models: Implementation

In addition to the LSTM model, more traditional ML models were also developed to create a framework for comparison of performance. LR and Linear SVMs were selected as models of this type since they are widely used for binary classification tasks. Additionally, since such models are not suited for sequentially-structured data, the cow data sequences developed for the neural network were converted into fixed-length data tables while indexing the respective train, validation, and test data splits by cow.

A behavioural sequence of length $T$ was transformed into a vector representation by explicitly including lagged information from previous time steps. Only features of the final observed hour, along with a few previous lagged observations, were extracted to provide information about the recent dynamic behaviour of the consumer before the prediction point. Apart from lag features, other statistical features about the sequence such as mean and standard deviation taken over the entire sequence window and taken into account to provide information about the total behavioural intensity of the sequence. Thus, by including both lag features and aggregations, the baseline models effectively get access to temporal information without necessarily modelling the sequences.

Both of the baseline models were trained on these tabular features using standardised features that were extracted only from the training data. LR produced probabilities directly, whereas the Linear SVM was equipped with a probability calibration layer for comparable probabilities. To provide a fair assessment despite the severe class imbalance, the threshold for each of the models was selected through F1-score optimization for the respective validation set, after which it was evaluated for the test set. Hence, a fair comparison is enabled between the models through equivalent windows of sequences, which allows for the illustration of the value of sequence models in ED.

# 5 Evaluation and Results

This chapter quantitatively analyses the effectiveness of the proposed ED framework and examines the factors that influence the effective application of such design considerations.

ED in dairy cows is a challenging task with inherent class imbalance, individual cow variability in behaviour, and the uncertainty of the onset of estrus. This implies that the performance of models developed for such tasks cannot rely completely upon an individual metric or point of aggregation. This chapter provides an assessment to the proposed a multi-perspective framework that considers performance on the population level modelling and individual level evaluation for various prediction windows.

This assessment follows a stepwise process. The chapter starts with the hyper parameter tuning results and the cross validation assessment in population level modelling, this is followed by the assessment of the LSTM model for its performance at a population level. Threshold-sensitive performance is also considered an essential component, given the high cost of both false alarms and missed estrus events in real-world farm settings. Along with these prformance analysis, the evaluation of baseline ML models is established for setting a performance benchmark for comparison purposes. In addition to its deterministic performance, this chapter explores prediction uncertainty via MC Dropout. By examining the correlation between uncertainty values and prediction accuracy, the results inform the model's prediction reliability. Lastly, we evaluate individual performance for each cow using individual-cow metrics to assess the model's generality with respect to individual cow behaviour.

## 5.1 Population Level Modelling

The performance of the proposed LSTM architecture at the population level is tested with four prediction intervals, namely current (0 h), 3 h, 6 h, and 12 h. For all windows, the proposed architecture provides probabilistic outputs where binary estrus predictions are made by selecting appropriate thresholds that maximise F1 scores of the validation set. The effect of threshold values on the proposed architecture is discussed in a dedicated section of this chapter.

### 5.1.1 Model Configuration and Setup

In this section, the empirical results of the model configuration and the validation procedure used to check the robustness of each configuration were presented. Although the previous chapter explained in detail the methodology behind hyperparameter tuning, along with model configurations and the training procedure, here the focus is on how different combinations of those configurations perform in practice and how stable they are. Hyperparameter tuning was conducted only on the population-level LSTM model for the prediction window of 6 hours. This was chosen as a compromise between the short-term detection and the longer-term prediction

and was selected to avoid biases towards the short term (current or 3h window) and the highly uncertain longer term (12h window). This optimisation setting was then used consistently for all the prediction windows and the individual level evaluations. Similarly, Cross-validation was performed solely at the population level, given that the model's individual level evaluation relies on cow-specific temporal splits rather than interchangeability between groups of subjects.

### Hyperparameter Tuning Results

The hyperparameter tuning for LSTM network is performed to find a robust set that can address the highly imbalanced nature of the problem in detecting estrus. Instead of presenting only the final optimal set that was selected, the best hyperparameters by ranking are highlighted in Table 5.1 below.

Table 5.1: Top five Hyperparameter Configurations Evaluated Through LSTM Model Tuning

| Rank | Batch Size | Dropout | Learning Rate | LSTM Units | Optimiser | PR-AUC | ROC-AUC | Best F1 |
|---|---|---|---|---|---|---|---|---|
| 1 | 32 | 0.4 | 0.0001 | 64 | Adam | 0.603 | 0.934 | 0.623 |
| 2 | 16 | 0.4 | 0.0010 | 32 | Adam | 0.591 | 0.937 | 0.612 |
| 3 | 16 | 0.2 | 0.0001 | 64 | Adam | 0.586 | 0.937 | 0.595 |
| 4 | 32 | 0.4 | 0.0010 | 32 | Adam | 0.585 | 0.936 | 0.603 |
| 5 | 16 | 0.4 | 0.0010 | 64 | RMSProp | 0.585 | 0.935 | 0.609 |

Hyperparameter tuning was performed by means of a structured grid search over a predefined set of hyperparameters. The hyperparameters that were explored included batch size 16, 32, dropout rate 0.2, 0.4, learning rate 1e−3, 1e−4, number of LSTM units 32, 64, and optimize r Adam, RMSProp. As a result, 32 different hyperparameters were explored, each of which was then tested on the validation set based on F1 and PR-AUC scores. The Best F1 score reported is the best attainable value of F1 on the validation set through threshold optimization on the predicted probabilities, rather than evaluating on a fixed threshold of 0.5. For all the hyperparameters tested, the ROC-AUC values ranged from about 0.925 to 0.937, reflecting that the ranking performance is highly stable against the hyperparameters. However, variation can be observed in the values of F1-scores, which is sensible to the choice of hyperparameters such as the learning rate, dropout rate, number of units in the LSTM layers, and the choice of the optimiser. The combination of dropout rates with adaptive optimisers like Adam or RMSProp resulted in better trade off between precision and recall. Both Adam and RMSProp were used during hyperparameter tuning, with the final hyperparameters chosen based on their validation performance.

To provide guidance on the selection of the model, an integrated tuning approach was employed. In terms of farm management, the precision involves the percentage of the predicted events of estrus that correspond to actual events, wheareas the recall involves the ability to detect actual events of estrus, which is important in the prevention of lost opportunities for insemination. A major focus was given on PR-AUC ( average precision), because it gives more

realistic performance estimates regarding the handling of extreme class imbalance, whereas ROC-AUC served as the secondary method to ensure the overall ranking consistency, and the F1-score calculated at the optimal threshold but not as an objective on its own.

Figure 5.1 shows the PR curve for the best hyperparameters on the validation set. This curve shows a stable precision level for a wide range of recall values and a significant improvement over the random classifier, representing a good ability to discriminate the minority class. The ROC curve represented by figure 5.2 shows a high level of separability which confirms the quality of the extracted features, although it is acknowledged that ROC curve-based evaluation can be overly optimistic in imbalanced problems.

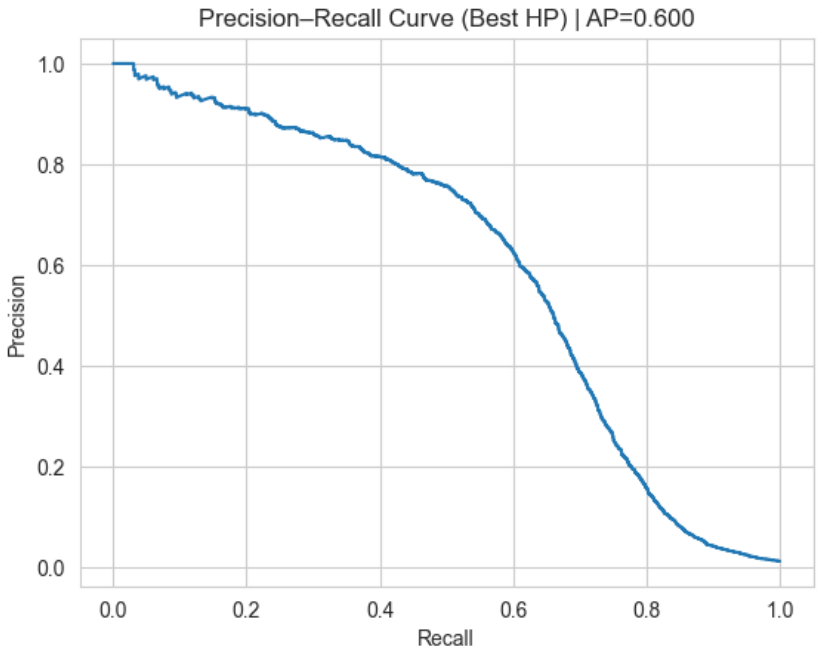


Figure 5.1: PR Curve Of Best Hyperparameter Set

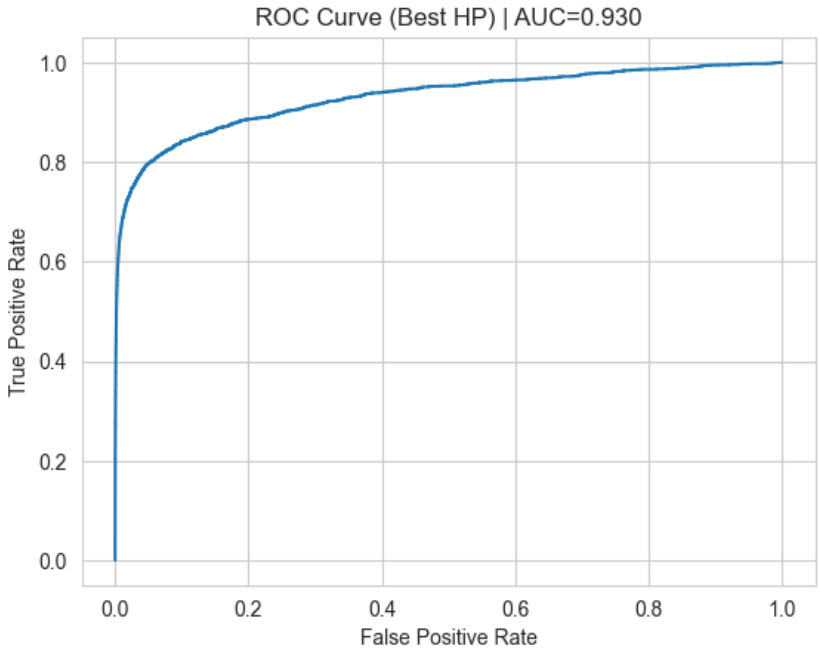


Figure 5.2: ROC Curve Of Best Hyperparameter Set

Only minor variations in performance can be observed for the top ranked parameters, which suggests robustness in the hyperparameter choices for the architecture for the LSTM model. The final settings for the model was selected on the basis of the PR-AUC value and used ROC-AUC as a tie breaker. This value has been subsequently used for all population level and individual level validations. Hyperparameter tuning was implemented at the population level modelling on the 6-hour prediction window, chosen as a representative point that will balance early predictions and the stability of behavioural signals. While in theory marginally better performance may be attained through individual window hyperparameter tuning, the same hyperparameter setting was used across all time horizons and individual evaluations to maintain consistency in methodological comparison. This is acknowledged as a limitation to the hyperparameter tuning process.

### Cross Validation Performance

To further assess the suitability of this chosen configuration, 5-fold cross-validation grouping by cow was performed. for every fold, the cows were split such that no individual cows appeared in both training and validation sets to avoid data leakage. This assesses the model using average results and standard deviation, which is highlighted in Table 5.2. This analysis evaluates model performance across different partitions of a given dataset, rather than a single split used in train-test separation. The validation is completed using cow based grouping in each folds which ensures that no individual cow appears in both train and validation folds. Cross-validation was only done for the population model in order to test its robustness and how well it generalized on different data splits. Cross-validation on the individual-cow model is not

meaningful because individual model already uses splits based on time for each individual cow and focused to analyse cow specific variability and consistency.

Table 5.2: Cross-Validation Performance Across Prediction Horizons

| **Prediction Window** | **PR-AUC (Mean ± SD)** | **ROC-AUC (Mean ± SD)** | **F1 (Mean ± SD)** | **Folds** |
|---|---|---|---|---|
| Current | $0.580 \pm 0.036$ | $0.981 \pm 0.005$ | $0.593 \pm 0.031$ | 5 |
| 3h | $0.537 \pm 0.087$ | $0.960 \pm 0.006$ | $0.560 \pm 0.052$ | 5 |
| 6h | $0.482 \pm 0.038$ | $0.928 \pm 0.007$ | $0.522 \pm 0.036$ | 5 |
| 12h | $0.349 \pm 0.115$ | $0.853 \pm 0.006$ | $0.434 \pm 0.072$ | 5 |

The performance of AUC shows relative stability with small standard deviations across folds for all prediction windows, indicating good ranking performance. With the prediction horizon increasing from the present detection to 12 hours, a gradual decrease in performance is observed. This is expected because, in earlier predictions, the model has to detect pre-estrus behaviour with less distinct patterns.

Consistently, the F1-score is more variable across folds, especially when considering a larger prediction horizon. F1-score variability can be attributed to class imbalance and to the unequal distribution of estrus events across splits. Such a characteristic is expected in a rare-event TS classification task, where this variation does not indicates model instability but rather indicate improved identification of such events in a TS.

Figure 5.3 also shows the cross validated PR-AUC values for different prediction windows, where each point depicts the mean PR-AUC value over 5 folds. The vertical lines shows one standard deviation of PR-AUC across different folds, which reflects the variations due to the change in train-validation splits rather than uncertainty within folds. The increasing width of the bars at 12h prediction window corresponds to growing variation across folds, which is expected in rare event prediction as the estrus cycles become less identifiable for distant prediction windows.

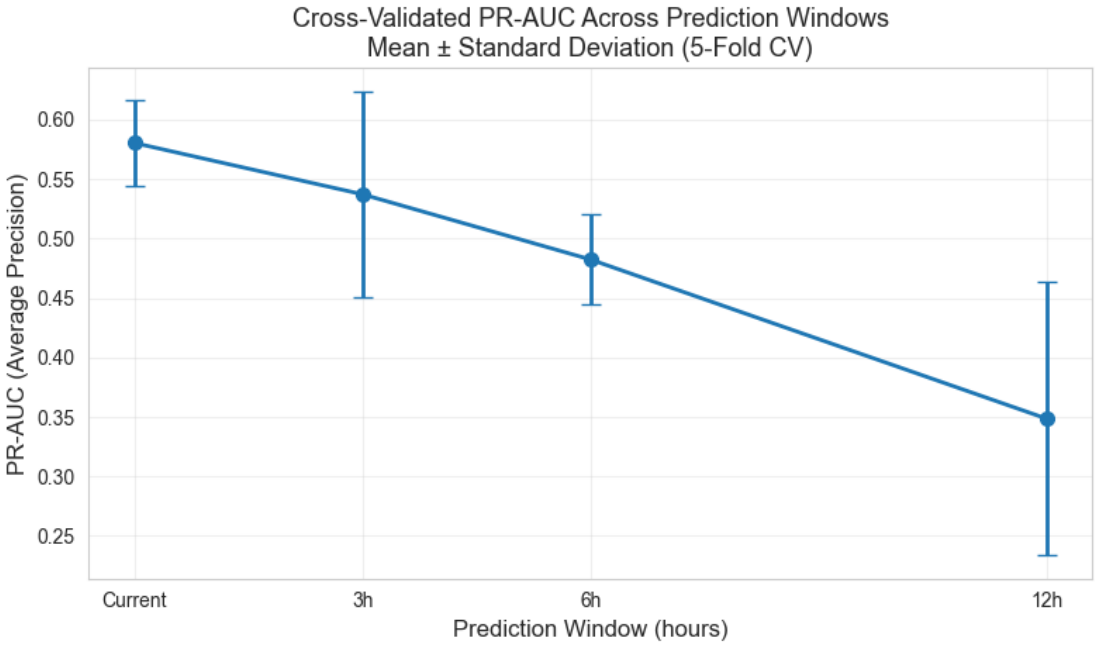


Figure 5.3: Cross validation : PR-AUC Curve

In summary, cross-validation shows that the chosen model configuration achieves consistent generalisation. The observed standard deviations confirm the model's robustness across different data splits. The standard deviation of all mean values validates model performance, ensuring it is not overfit to a particular split and providing a sound basis for analysis at the population and individual levels. Based on the configuration and validation results above, the next sections thoroughly analyse performance. First, we present population-level analyses, including baseline comparisons, followed by analyses of thresholds, uncertainty, and individual- and cow-level analyses.

### 5.1.2 Population Level LSTM Performance

Figure 5.4 showcases the population-level performance of the LSTM model on various metrics in different prediction windows. The values of this metrics are shown using a graphical representation instead of a table format to better identify the relative variation of the scores. The accuracy values remain high irrespective of the prediction window. This can be expected in an ED task because of the dominance of non-estrus instances in the dataset. Hence, AUC is seen to be the main metric to evaluate the discrimination ability of the population level rather than the accuracy metric because it represents the ability to distinguish between estrus and non-estrus regardless of decision thresholds and class imbalance.

The AUC values show high discriminative power for short-term forecasting intervals, especially for the current and 3-hour windows. As the forecast window extends, a steady decrease in AUC is noticed and it indicates the weaker predictive signal further from estrus period.

The variation of the precision, recall, and F1 scores for the proposed LSTM model on all proposed prediction windows are also showcased in the figure 5.4. For smaller forecast windows, recall scores are high which reflects a strong response to estrus events closer to onset. Precision shows a relatively stable or mildly rising pattern with larger windows, due to more conservative forecast decisions at higher thresholds. On the other hand, as the forecast interval is extended, recall measures drop, mirroring lower detectability of estrus-driven behavioral patterns at longer forecast distances.

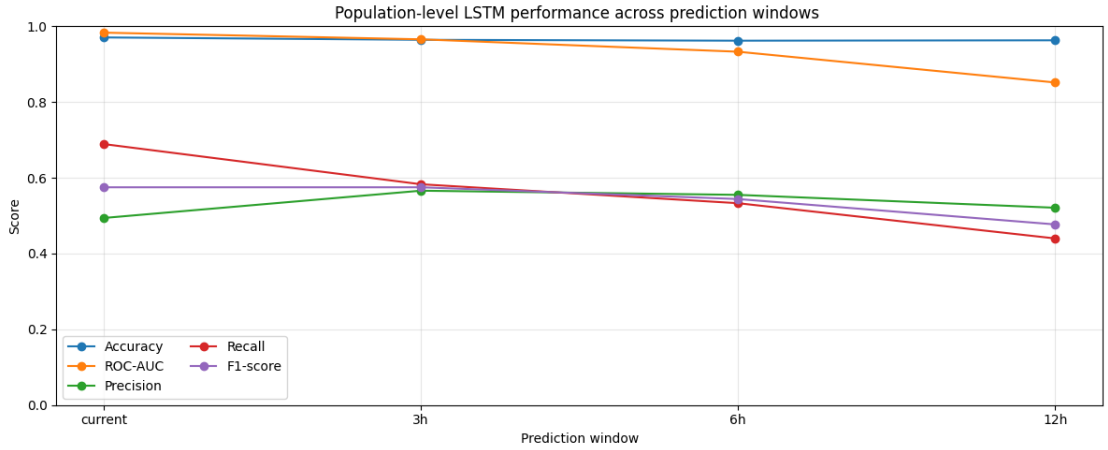


Figure 5.4: Population Level LSTM Performance Across Prediction Windows

Figure 5.5 illustrates the positive class rate (pos_rate) in the testing data for various prediction windows for the population level. The positive class rate is calculated as the ratio of estrus instances to the total number of test sequences. The positive rate is always below 1.5% which confirms the presence of extreme class imbalance. To ease understanding, note that a positive rate of 0.01 actually means 1% estrus positive in the test data.

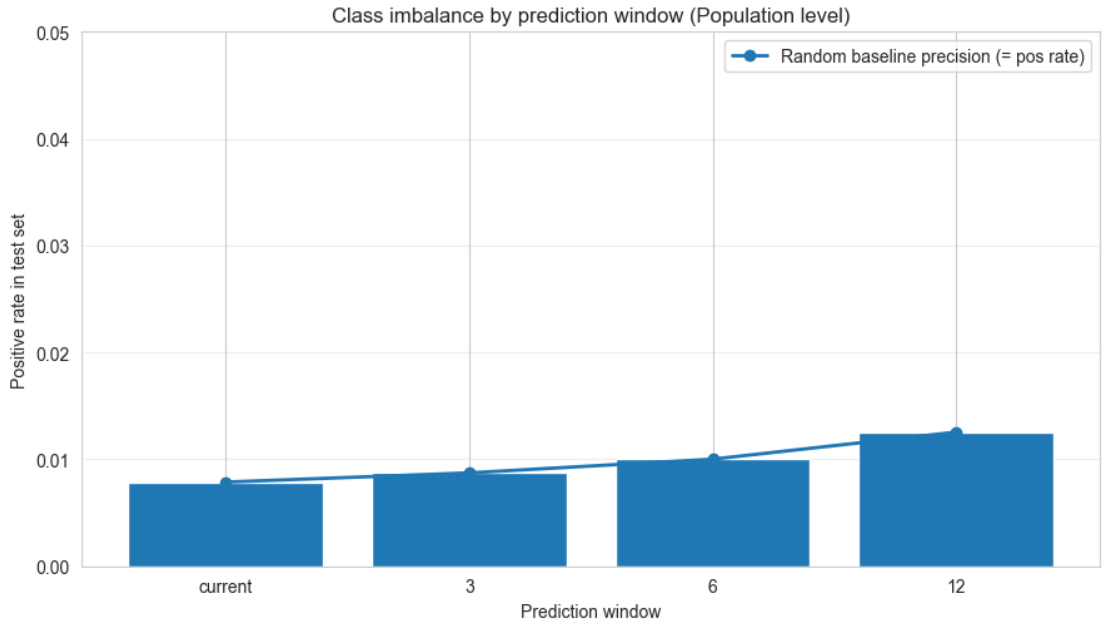


Figure 5.5: Prevalence of Positive Class by Prediction Window

This increase in the positive rate for larger windows can be expected because larger windows have higher chances of at least an estrus event falling within the window. However, it can be noticed that the distribution of the dataset dominantly inclined towards the negative class for all windows. This imbalance has high influence in the performance of threshold-dependent metrics. The metrics such as precision, recall, and F1 measures are highly sensitive to the prevalence of the positive class, and hence their performance should be evaluated in the context of class distribution. The F1 score decreases monotonically with the prediction window length. The reason for this is the joint contribution of the decrease in recall and the persisting class imbalance problem, rather than any degradation in the intrinsic discrimination power of the model itself , already accounted for in terms of the AUC scores reported earlier. The model achieved AUPRC values ranged between 0.583 for current window to 0.406 for 12 hour window.

In summary, the results of population-level testing reveal that the LSTM network maintains high levels of discrimination ability over all prediction intervals, with a particular emphasis on short-interval estrus recognition. The combination of providing threshold-independent measures of accuracy and AUC coupled with graphical displays of class imbalance issues reveals a well-rounded understanding of model performance.

### 5.1.3 Threshold Analysis

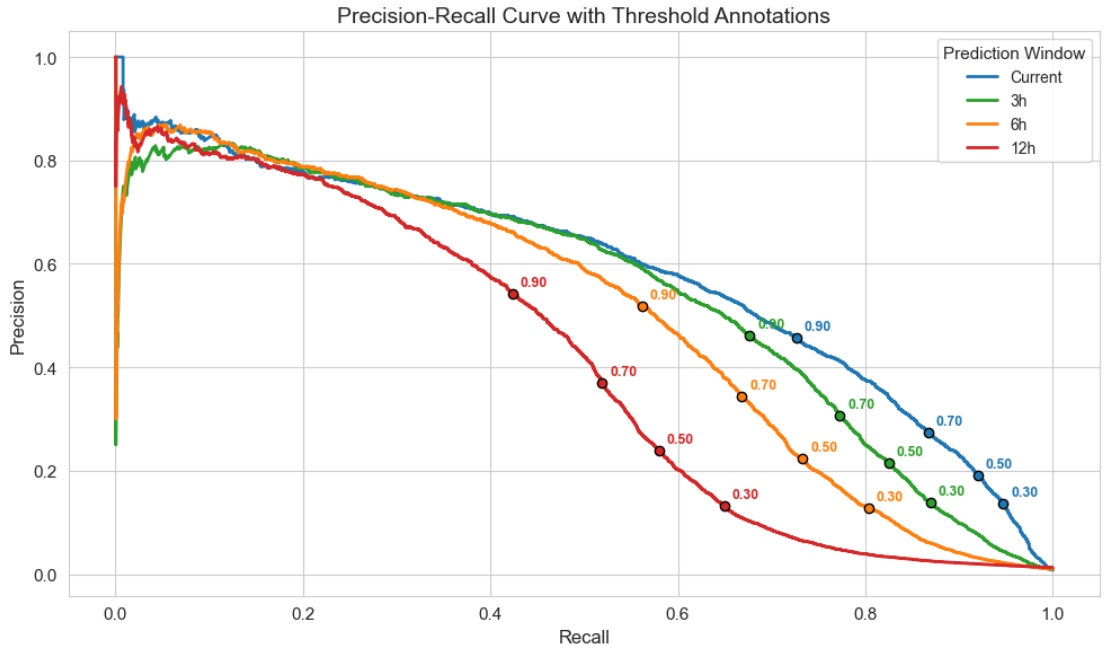


Figure 5.6: Precision-Recall Curve

The LSTM outputs probabilities for estrus occurrences, whereas, the conversion of probabilities into binary classes is done by setting an appropriate threshold value. However, due to high class imbalance in estrus prediction, a value of 0.5 as a threshold value may not be accurate and can lead to high false positives or missed occurrences of estrus. Hence, decision threshold values were adjusted individually for each prediction horizon using the validation set, with the aim of maximizing F1-score values. The model achieved a F1 score of 0.57 at the current window and decreased to 0.47 at 12 hour window.

In Figure 5.6 represents the PR curve for the population-level model using LSTM across various prediction windows. The x-axis represents the recall values, and the y-axis represents the precision values, while the points marked with different colours throughout the PR curve represent various thresholds within the model.

The noticeable fluctuation near the origin of the recall axis is associated with the highest possible threshold value, and in these areas, only a few positive samples are predicted as positive due to the strict decision threshold. The reason for the peak in the initial part of the PR graph is the fact that precision becomes extremely sensitive when the number of positive examples predicted is very low. In the case of the current prediction window(0h), the PR curve indicates that the precision level remains relatively high for a wide range of recall values. Some false positives may also point out the errors in the human based labels provided, as high activity events unrelated to estrus because of relocation or environmental factors can resemble estrus related activity fluctuation and possible to be overlooked by experts (information provided by

Farmtech a.s.). Based on the PR curve, the best operating point was found at a threshold of 0.95, which corresponds to the point which indicates maximum F1 value, whereas this values are illustriated in Figure 5.4. Despite being close to 1.0, the chosen threshold corresponds well with the underlying class imbalance problem and indicates that the system predicts true events of estrus with a high degree of confidence, which prevents the false alarms in practice. With the extension of the prediction horizon, the PR curves demonstrate a continuously steeper drop, which indicates a growing uncertainty of predictions when they extend over a longer period of time.

The results indicate that the selection of threshold is strongly dependent on windows and a fixed threshold across all windows is not suitable. The findings of the PR study are a testament that proper threshold adjustment in the pursuit of optimizing rare event detection classifications remains critical in making appropriate trade-offs between the numbers of false positives and false negatives. Hence, the selected thresholds were used uniformly in subsequent evaluations.

### 5.1.4 Baseline Comparison

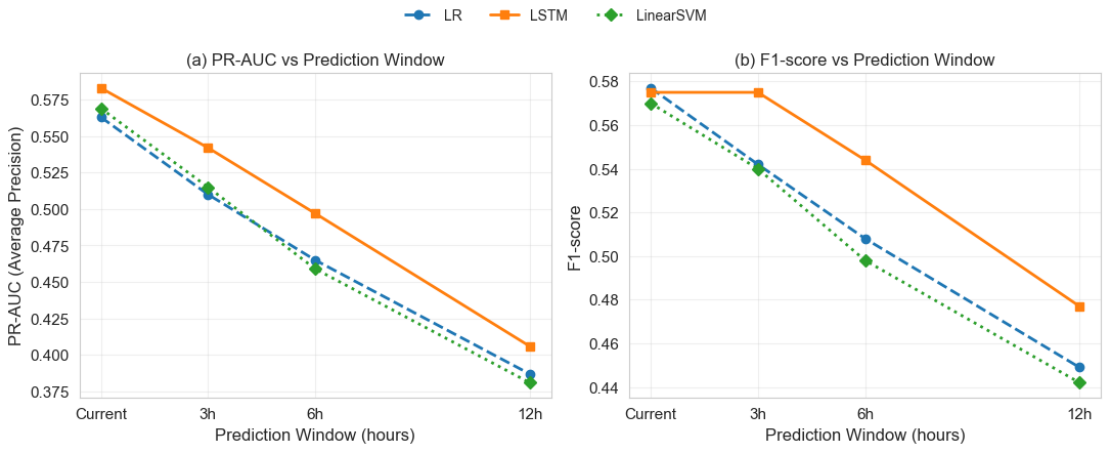


Figure 5.7: Baseline Models Comparison With LSTM

To evaluate the performance of the proposed LSTM model, results at the population level were compared with two traditional baseline classifiers, namely LR and Linear SVM. All three models were tested with the same set of input features, data splits, prediction windows and hyperparameters with an aim to ensure fairness and consistency. Decision thresholds for all three models and time windows were determined through validation set optimization of the respective F1-scores, allowing each model to make decisions at their most balanced point considering the highly imbalanced nature of their classes.

Figure 5.7(a) plots PR-AUC for all three models against prediction windows. PR-AUC indicates each model's ability to predict estrus-positive instances before estrus-negative instances given different operating thresholds. For all prediction windows, the LSTM model has the maximum PR-AUC values, which signifies its better discrimination power than LR and Linear SVM.

For all the prediction windows in the future (3h to 12h), the LSTM always gains approximately 1.5-3.2% in PR-AUC over the LR and Linear SVM. For smaller prediction windows, there is a larger gap in PR-AUC values between the three models because behavioural observations are relatively close to the estrus event in the sequences and are therefore most representative. For larger prediction windows, PR-AUC values drop for all three models because estrus prediction becomes more difficult as the prediction window increases.

Figure 5.7(b) shows related F1-score patterns. F1-scores quantify a trade-off between precision and recall for a particular operating threshold and are suited to rare event detection. The LSTM model is competitive with other models in terms of F1-scores for all prediction windows. The LSTM reflects absolute improvements in F1 score of approximately 2.8 to 3.6% points for prediction horizons larger than 3 hours reveal a better PR trade-off of future estrus predictions. The LSTM model achieves a higher F1 score at 3h, 6h and 12h, while the f1 score at current window is comparable across all models. For time windows beyond 3 hours, all models exhibiting a degradation in performance, however, the proposed LSTM technique achieves slightly higher f1 score compared to the baseline models.

The relatively small performance difference between the performance of the LSTM approach and the baseline models is expected and does not reflect a weaker performance of the approach presented in this study. All models were trained on the exact same set of engineered behavioural features, which contain a lot of information about the estrus cycle. The linear models are generally able to learn about the signal, although the LSTM approach is a much better model of such a signal due to the temporal modelling and the ability to learn non-linear dependencies.

A set of meaningful baselines has been established here to serve as a basis for further evaluations in the context of the LSTM model. Although this displays the reasonable performance at the population level of the classical ML methods, there is a point at which the deficiencies are increasingly apparent once the forecasting window and temporal complexity are factored inwards. Thus the baseline comparisons not highlighting the large performance gaps, but contextualising the added value in the sequence based modelling approach.The incremental improvements which can be observed in the highly imbalanced ED, where smaller gains in PR-AUC and F1 score indicates improved reliability in decision making.

### 5.1.5 Uncertainty Analysis

In order to measure predictive uncertainty in ED, MC Dropout was used to apply dropout during prediction in the population level LSTM network. The approach allows multiple stochastic forward passes through the network by enabling the dropout. It provides probabilities of prediction for each sample. The standard deviation of these predictions indicates uncertainty, representing how confident the model is in its outputs.

Figure 5.8 below highlights the comparison in predictive uncertainty for correctly versus incorrectly classified instances for all prediction windows. In all prediction intervals, the uncertainty in the incorrect predictions is shown to always be higher than that of the correct predictions.

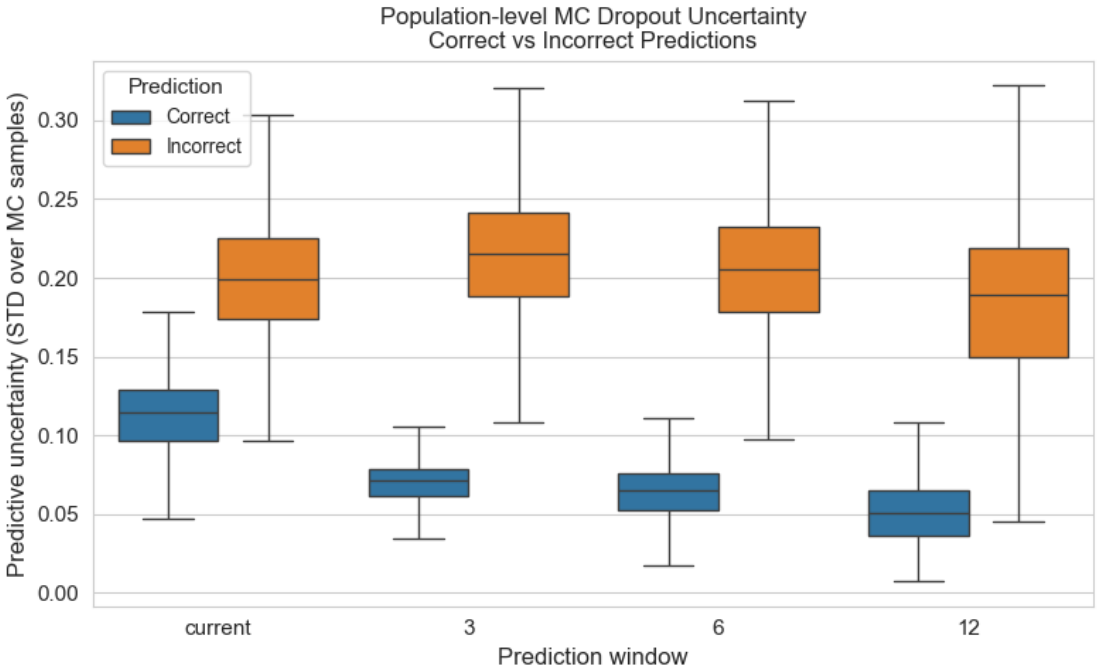


Figure 5.8: Uncertainty Standard Deviation Of Correct and Incorrect predictions

These result illustrate that there is a strong connection between prediction uncertainty using MC Dropout and prediction reliability, such that predictions with low levels of uncertainty are more likely to be correct, while predictions with higher levels of uncertainty are more likely to be cases of misclassification. This is highly desirable for systems designed for safety-critical applications as well as decision support.

In addition, the mean and median uncertainty measures of correct and incorrect predictions affirm the above-discussed separation. To illustrate, on the current prediction window, the median value of uncertainty for correct predictions is 0.114, while the median value for incorrect predictions is 0.199. A similar disparity is noticed on the 3 h, 6 h, and 12 h prediction intervals.

**Statistical Validation of Uncertainty Separation :**

To validate the observations statistically, a non-parametric Mann-Whitney U tests were implemented for every prediction window. The purpose was to verify if uncertainty is increased for incorrect predictions compared to correct predictions. In each situation, the null hypothesis was rejected with a $p$ value less than 0.001 across all prediction windows, which indicates that there is a significant difference between both groups.

In addition, Spearman rank correlations were carried out, which indicated a significant negative correlation between the frequency of correct predictions and the amount of uncertainty for all tested horizons($p < 0.001$). The correlation values varied between -0.216 and -0.270. This specifically implies increased uncertainty is asscociated with the higher probability of wrong predictions, which further confirms the reliability of the uncertainty values.

Taken altogether, it is evident from this uncertainty analysis that MC Dropout can obtain informative uncertainty estimates for population-level estrus prediction tasks. The uncertainty levels are varying in accordance with the prediction window and are able to effectively differentiate correct from incorrect predictions. This further validates that uncertainty estimates can serve as an additional performance criterion together with other performance measures to facilitate more conservative and well informed decisions in automated ED systems.

## 5.2 Individual Level Modelling

### 5.2.1 Individual Level LSTM Performance

After the evaluation on the population level, the performance of the proposed LSTM model is investigated under an individual level split, where the data split were based on each cow using the temporal split, which means about 70% of each cow's sequences to the training set, 10% to the validation set, and the remaining 20% to the testing set. The sequences build this way were pooled across cows for training single model, while preserving the temporality. Although the same dataset is used, the evaluation method was modified in order to assess temporal prediction consistency within cows temporally. The test set is set to follow the training set, thereby avoiding the leakage of information by considering the fact that predictions of estrus occurrence was performed based on past observations in actual deployment scenario.

However, the existence of the class imbalance observed at population level also presents in the individual level, given that both assessments are evaluated from the same dataset. Nevertheless, due to the cow-specific splitting within time and limited availability of cow data, there could be a slight variation in the actual composition of the testing set for each cow. In particular, there are a few cows for which there are no positive estrus observations within the testing set, and these have a direct impact on the metrics of precision, recall, and F1 measure at the cow level.

Figure 5.9 shows that the LSTM model maintains a good capability for distinguishing estrus from non-estrus within temporal splits. Within the current prediction window, it reaches up to a high AUC value, representing a clear separation of estrus and non-estrus cases even when the classes are severely imbalanced. The result is quite consistent with the biological fact of estrus behavior which symbolises the fading of behavioral signs as going farther from the onset.

The model performance in terms of precision, recall, and F1-score is also captured in figure 5.9. Also consistent with the above results, the performance tends to reduce as the horizon increases. Performance peaks when the horizon was either the current time or the 3-hour horizon. Beyond that, there's a steady reduction as the horizon increases which is likely due to the rise in the uncertainty of the timing of the onset of estrus.

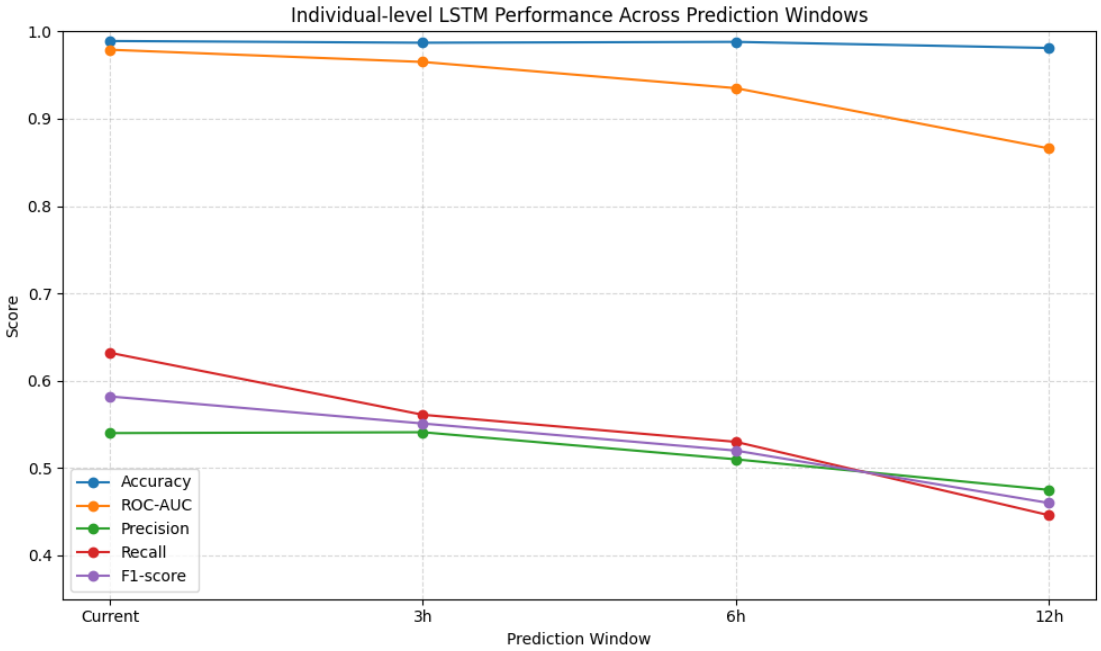


Figure 5.9: Individual Level LSTM Performance Across Prediction Windows

The results on the individual level show that the LSTM approach are stable, even under extreme conditions of class imbalance and varying availability of test samples per cow. The model achieved AUPRC values ranged between 0.552 for current window to 0.367 for 12 hour window, also indicates the increased difficulty in higher prediction windows. The results complement the population level analysis and prove the effectiveness of the proposed method under complementary within-cow validation setting.

### 5.2.2 Threshold Analysis

The presence of class imbalance and the individual basis development, the optimization of the decision threshold was done separately for each forecasting window. Figure 5.10 shows the PR curves with the corresponding decision thresholds for the individual level model.

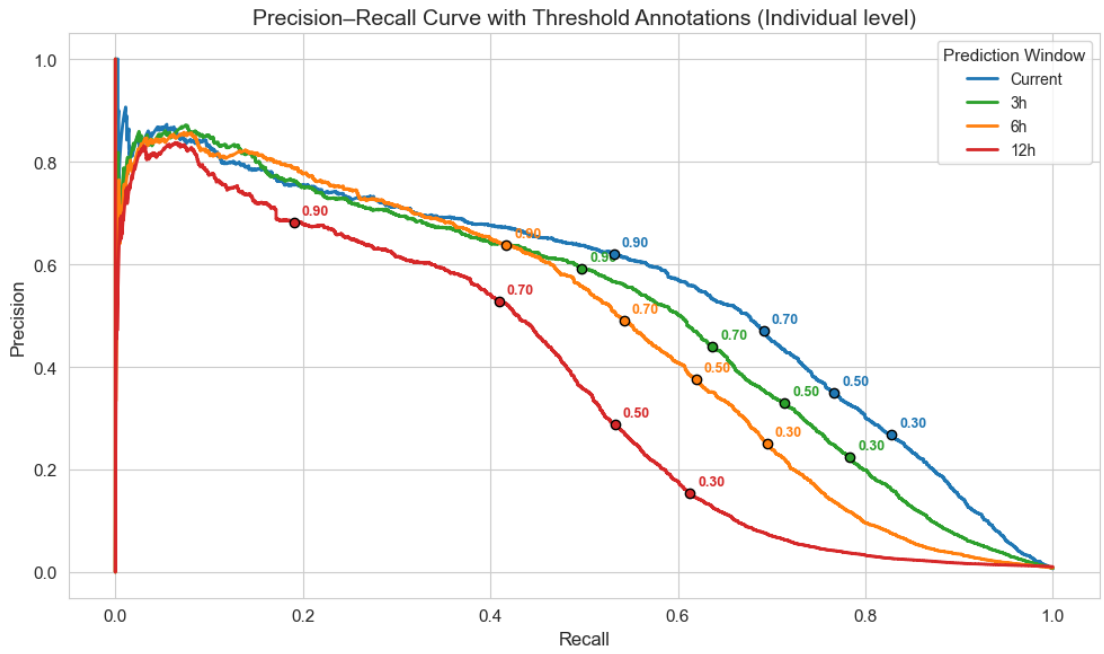


Figure 5.10: Individual Level Precision-Recall Curve

The PR curves illustrate the expected trade-off between precision and recall based on the threshold values. When considering the individual level data and the shorter horizons (current and 3 hours), the PR curves indicate that there is a smooth drop-off of precision with respect to an increase in recall, and this reflects the separability of the estrus and non-estrus samples. However, with the progression of the prediction horizon to 6 hours and then to 12 hours, there is a sharp drop-off, and this reflects the decreased signal strength far from the onset of estrus.

One thing that is evident in the PR graphs is the fluctuation in the beginning, especially in the recall values. It is attributed by two aspects that are embedded in this setting involving individual data. First, events in estrus are very few in relation to the data per cow, which basically translates to the point that the first predicted values are for a few data points with very high confidence. As a result, small movements in the curve could lead to very large changes in precision due to these confidence levels. Second, the range associated with very high threshold values made the predictions made through leveraging a few sequences with high probability values. such changes are common in rare event modelling and do not indicates the model instability.

Annotations along the curves represents the threshold values and illustrate the trade offs between precision and recall as functions of operating points. Higher thresholds bias toward precision by considering only predictions with high confidence in estrus, whereas lower thresholds bias toward recall, with higher numbers of false positives. In summary, from the analysis of the PR curve, it can be confirmed that individual performance is very sensitive to threshold values. The results are consistent with biological aspects of estrus development and aligns with statistical properties of highly skewed and cattle specific datasets. The threshold values

determined is applied to all prediction windows when reporting individual level evaluation and baseline comparison.

### 5.2.3 Baseline Comparison

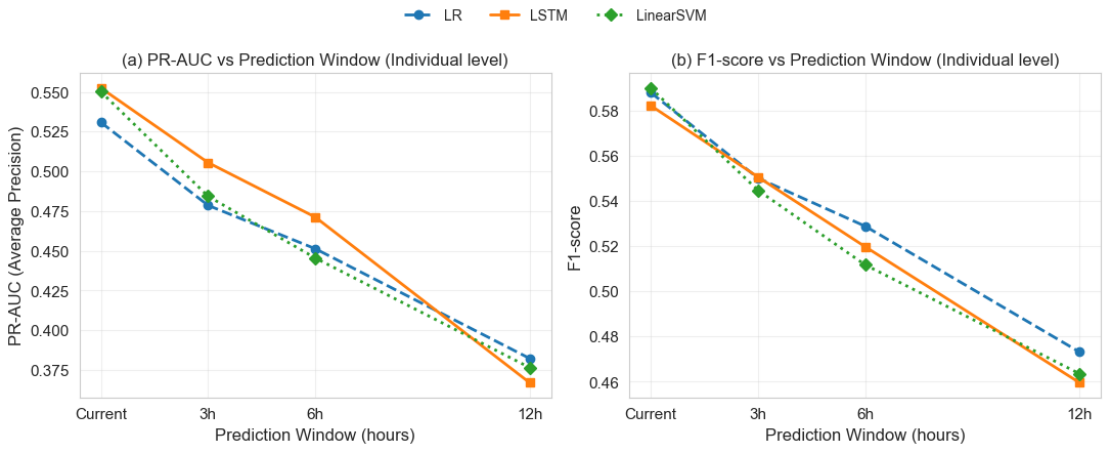


Figure 5.11: Baseline Models Comparison With Individual Level LSTM

Figure 5.11 presents a visual representation of comparison on both PR-AUC and F1 metrics for different prediction windows. The choice of evaluation metrics was made to reflect the strong class imbalance known to exist within tasks of ED.

As seen in Figure 5.11(a), all three models demonstrate a gradual decrease in PR AUC with the increase of the prediction window. This behavior can be expected because larger prediction windows imply a lesser degree of temporal proximity to estrus onsets and thus a weaker strength of discriminative signals. In all the windows, it is evident that PR-AUC values for the LSTM model are comparable to or slightly better than those of the competing models, specifically for both the current and 3-hour window. This confirms that the LSTM model is capable of capturing temporal dependencies in individual cows.

However, at longer time-horizons (6 hours and 12 hours), the PR-AUC scores for all models tend to a common value, signifying a difficulty level in the prediction task irrespective of model complexity. Most importantly, there is no drastic degradation of the LSTM model.

In Figure 5.11(b), the trends of the F1-scores for the prediction windows are illustrated. As in the PR-AUC scores, there is a monotonically decreasing trend for all models with the increasing size of the horizon length, due to the inherent challenge in making long-term estrus forecasts for individuals.

The F1-scores of the LSTM overlap tightly with LR and Linear SVM on every horizon, showing only a small lead on shorter intervals. This result suggests that while traditional models perform well on engineered lag based features, the LSTM model is capable of strong performance on the sequential data under pooled individual level training. This implicates that the selection of LSTM is not on the basis of raw F1 alone, but model capability and intrinsic reliability. These comparable levels of model performance on the individual level reinforce two

important considerations of the task. Firstly, the task of individual-cow forecasting tends to be more noisy than modeling on the population level of the herd because it involves different patterns of behavior and relatively few instances of desirable outcomes. Secondly, the lack of a large performance gap in the task does not make the use of the LSTM model less valuable, rather it shows that the model performs comparable performance along with the temporal modelling flexibility and suitability to the uncertainty estimations.

### 5.2.4 Uncertainty Analysis

To evaluate the level of predictive uncertainty for individual cows, MC Dropout was used during test for individual LSTM models. Following through with the pattern established during population-level analysis, the dropout layers were engaged during prediction which allows for multiple stochastic forward passes for individual test sequences. The standard deviation of probability estimates for MC samples was taken as the estimate of predictive uncertainty, where larger values correspond to lower confidence in model estimates.

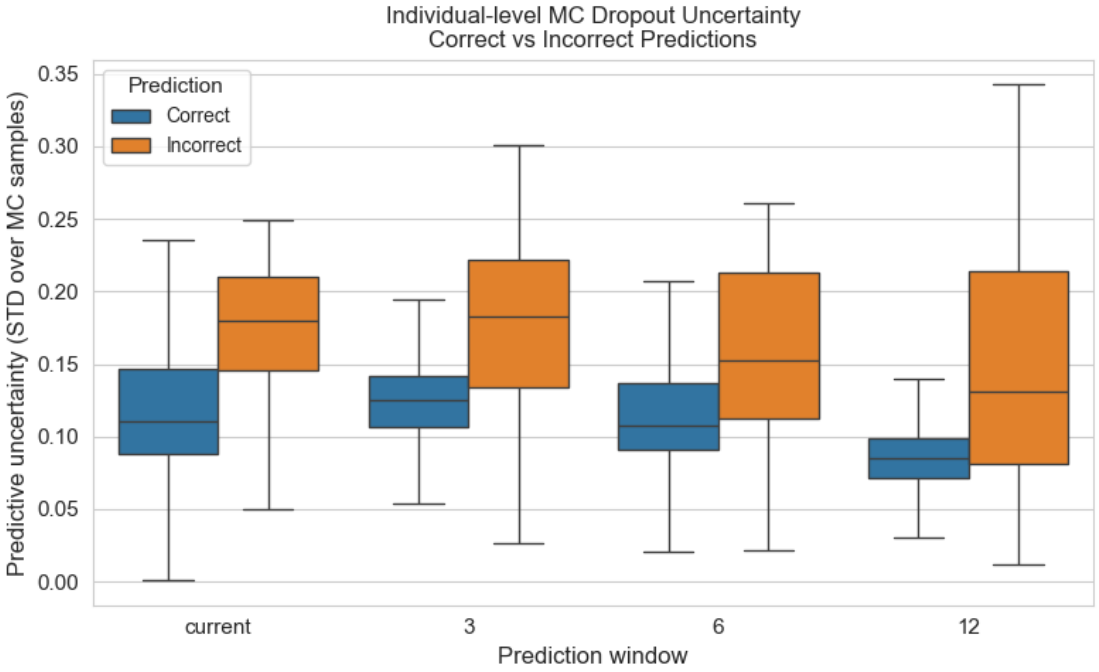


Figure 5.12: Individual Level - Uncertainty Standard Deviation Of Correct and Incorrect predictions

Figure 5.12 shows a contrast of predictive uncertainty between correctly and incorrectly classified data points across all time windows. For all time windows combined, a higher median uncertainty correlates with incorrect predictions than with correct predictions. This points out that the individual-level predictions can also be distinguished based on uncertainty. Considering the findings at the population level, there is a partial overlap of the two distributions. This

can be attributed to the fact that individual level models can be difficult to work with since a weak or irregular estrus signal can correspond to a moderate level of uncertainty, even among correct predictions. Despite the overlap, a systematic change in central tendency is observed in all cases, where the wrong predictions always present higher levels of uncertainty on average. This points to the existence of relative uncertainty separation at the individual level in the task of rare event detection of different individuals using only a few positive cases.

**Statistical Validation of Uncertainty Separation :**

In each case, the null hypothesis that the distributions of uncertainty values for correct versus incorrect predictions was rejected with a $p$ value less than 0.001 across all prediction windows. This means that the results are statistically significant although there is some overlap. Moreover, Spearman rank correlation analysis revealed that there was a valid negative relationship between the correctness of the prediction and predictive uncertainty for all time windows.The correlation coefficients ranged from $-0.073$ to $-0.091$ across prediction windows. The magnitude of the correlations is lower compared to population level, whereas all those values are uniformly negative and statistically significant. The correlations were lower in absolute terms than those found at the aggregated level, it confirms that greater uncertainty is associated with the chances of misclassification.

Considering everything, these results show that MC Dropout provides good and well calibrated analysis of uncertainties at the individual level for ED. Although the uncertainties are not as separable as at the population level, the estimates are still informative and accurate, especially with respect to detecting predictions with low confidence values.

### 5.2.5 Cow level Performance Analysis

Individual level analysis enables a complementary evaluation of the model performance within individual cows. Per cow evaluation is most applicable to the analysis of ED because the behavioural patterns vary significantly between the cows. The cow level assessment remained only within the framework of the individual level model since the population level model focuses on the generalisation of the unseen cows and does not entail per-cow performance assessment.

The population-level approaches have been left out of the comparison as their data allocation method puts individual cows entirely into either the training or testing data groups. It makes per-cow measures less informative and hard to compare since each testing cow contributes to its own sequences and is not partially seen during the training process. Accordingly, only the individual-level approach allows for meaningful per cow evaluations.

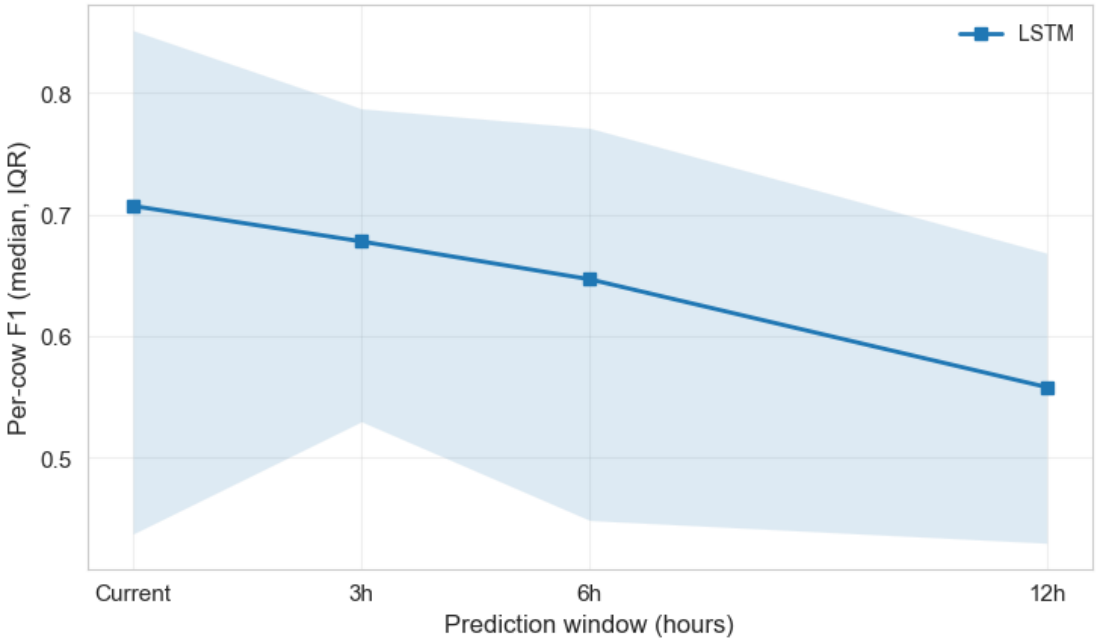


Figure 5.13: Cow Level Performance Analysis

In Figure 5.13, the median value for the per-cow F1 score for each prediction window is shown, and the Interquartile Range (IQR) shown as shaded band. It gives an estimate of how the performance potentially range. The median value of the per cow F1 score reduces steadily as the window size for the prediction of estrus stages advances, which logically matches the increasing uncertainty involved with longer prediction intervals. However, the IQR remains fairly consistent.

The consistent decline in the median per cow F1 score without sudden changes in the IQR indicates that there is smooth degradation of performance on individual instances, rather than sudden failures. Although the lower quartile showcases mild shifts due to the increase in availability of positive sample in higher windows, the decrease in median of f1 points out the increasing difficulty in predictions. This is a highly desirable trait when deploying models, suggesting that it is robust to individual differences while maintaining biologically realistic performance paths. Even though there is a high level of variability when displaying per-cow distributions of F1 scores due to a significant imbalance between classes at the herd level, using median and IQR aggregation provides a reliable and insightful indicator of this performance.

The performance across individual cows showcases significant variability which also effects the ED. Several causes contributes the exhibition of such behavior:

- The rare estrus events causing uneven positive samples.
- Cows show individual variations in activity level, intensity of estrus, and quality of sensor signals.

- Some cows have very few examples that test positively. That makes metrics like F1 score highly unstable and varying.

Therefore, it does not necessarily indicate failure of the model to have lower average per-cow values for the F1 metric. This phenomenon of having strong population level performance and poor individual performance metrics has been known to occur in the realm of imbalanced TS classification problems.

# 6 Discussion

This thesis deeply analysed the ED from TS data collected through wearable sensors, applying an LSTM approach on various estrus prediction windows (0 h, 3 h, 6 h, and 12 h). At the population level , the ROC-AUC values for the LSTM model were 0.983 (0h), 0.966 (3h), 0.933 (6h), and 0.852 (12h), while PR-AUC values dropped from 0.583 (0h) to 0.406 (12h). The individual level results for cow temporal split, ROC-AUC values varied from 0.979 for 0h to 0.866 for 12h, and PR-AUC values varied from 0.552 for 0h to 0.367 for 12h. At the population level, the analysis shows generalisation to unseen cows, while at the individual level, it demonstrates temporal prediction consistency within same cows. Together, this showcases that the proposed LSTM model is capable of generalising estrus patterns, thereby addressing the first research question. For both strategies, accuracy decreased with increasing timescale, consistent with real-world expectations. The time elapsed prior to estrus affects the strength and detectability of the behavioural signal.

As ED is an event which occurs rarely, and it points out the need for the analysis to be understood in light of the class imbalance, especially on an individual cow basis. This leads to threshold-dependent measures such as precision, recall, and F1-score also being highly dependent on this choice and changing dramatically even for marginal changes in the number of detected positives. This is especially evident in the initial portion of the PR curves for a scenario in which very few high-probability positives exist; a change in the threshold induces large swings in precision due to the addition or removal of predicted positives. Baseline comparisons show that the traditional tabular models remain competitive. This can be seen in longer time windows, where the discriminative information is weak for all models. However, the LSTM model performs comparably well on sequential representations and also provides an indicator of predictive uncertainty using MC dropout.

On future horizons ranging from 3 to 12 hours, the LSTM offered a steady advantage of about 1.5 to 3.2 percentage points in PR-AUC and 2.8 to 3.6 percentage points in F1 than LR or Linear SVM, showing the value of modeling sequence in forecasting. This has significant implications for decision support systems related to reproductive management, where not only the classification information but also its certainty, as indicated by measures of predictive uncertainty, are important. The results demonstrated that incorrect predictions generally have higher predictive uncertainties. At the level of the individual cow, the aggregated median and IQR trends imply a smooth degradation with increasing window size, without any particular points of instability. It indicated a degradation in performance, rather than a complete failure at various windows. Nevertheless, at the individual cow level, there is always a possibility of variation due to differences in the presentation of the estrus cycle, sensor noise, or the number of positive samples per cow. The demonstration of how practical applicability can be enhanced through the uncertainty estimations and window-specific thresholding, along with the degradation of performance with increasing prediction windows, answers the second research question.

In comparison with previous studies, the ROC-AUC values observed in this work are in line with more recent studies using LSTM architectures for ED, such as Chen et al.[47], who showed an AUC of about 0.95 for 24-hour sequences of behaviour. Previous studies using sensors showed high sensitivities (92% sensitivity and 89% specificity in Rutten et al. [41]), but often used heuristic methods and remained vulnerable to false-positive results in the presence of behavioural variables. Activity based methods (At-Taras and Spahr [36]) clearly improved upon visual observation, and this work follows in this line by providing an estimate of performance over multiple prediction horizons and providing uncertainty estimation for decision support.

Overall, the findings confirm the thesis that combining sequence modelling with window-specific thresholding and uncertainty estimation provides a strong, decision-relevant framework for ED, especially for nearby windows where behavioural signals are strongest.

# 7 Conclusion And Future Scopes

The thesis focused on addressing the challenge of automated ED of wearable sensor data, examining its predictive capabilities across various time intervals leading up to estrus incidence. Unlike other studies that focus on a specific point in system operation, this thesis considers predictive reliability and uncertainty across various time windows to estrus incidence, as enabling an early notification system requires a clear understanding of these aspects.

The results show a stable, information-rich signal for current window predictions, but a degradation in performance with longer windows due to higher behavioural variability and reduced physiological signals. It has been observed that using accuracy as a single metric is inadequate in this setting, and a horizon-dependent threshold and specific measures to address severe class imbalance are required. The observed consistent degradation of F1-score and PR-AUC with increasing window distance captures a fundamental difficulty of this particular problem, rather than any specific inadequacies of the modelling.

Another major contribution of this research study is the inclusion of MC dropout-based uncertainty estimation. The analysis of uncertainty has demonstrated a systematic relationship between the level of prediction confidence and the prediction windows, even when the boundaries between correct and incorrect predictions are not clearly defined in individual-level validations. Such an observation aligns with expectations for rare-event prediction in a heterogeneous group of subjects and validates the idea that uncertainty should be treated as a probabilistic rather than a deterministic quantity.

Future scopes of this study may utilise these results in several directions. First, the feasibility of aggregate predictions via a cascaded window-prediction approach can be examined, in which overlapping predictions for the target time window at the sequence level are combined into a robust final prediction. Second, RF could also be used as an additional baseline for comparison due to its non-linearity character and its ability to represent interactions between variables without having to model sequences. Lastly, the generalisability of the findings can be improved by including more farms and varying the observation time.

To conclude, this thesis shows that ED is a window-dependent classification problem, where its difficulty and optimal threshold points were strongly related with prediction horizons. Based on evaluations of performance, thresholds, and uncertainty on both population and individual levels, the thesis provides a solid foundation for a robust decision-support system in precision dairy farming.

# Bibliography


[1] M. J. VandeHaar and N. St-Pierre, "Major Advances in Nutrition: Relevance to the Sustainability of the Dairy Industry," *Journal of Dairy Science*, vol. 89, no. 4, pp. 1280–1291, 2006, ISSN: 0022-0302. DOI: 10.3168/jds.S0022-0302(06)72196-8. Accessed: Oct. 12, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022030206721968.

[2] *Dairy Herd Management Market is expected to generate a revenue of USD 5.28 Billion by 2031, Globally, at 6.80% CAGR: Verified Market Research*. Accessed: Dec. 7, 2025. [Online]. Available: https://www.finanznachrichten.de/nachrichten-2025-09/66500251-dairy-herd-management-market-is-expected-to-generate-a-revenue-of-usd-5-28-billion-by-2031-globally-at-6-80-cagr-verified-market-research-008.htm.

[3] Food and agricultural organisation of the united nations, *I1522e02*. Accessed: Oct. 12, 2025. [Online]. Available: https://www.fao.org/4/i1522e/i1522e02.pdf.

[4] J. Parish, "Estrus (Heat) Detection in Cattle," [Online]. Available: http://www.ext.msstate.edu/sites/default/files/publications/publications/p2610_0.pdf.

[5] *At what age is it safe to breed a heifer? | UNL Beef | Nebraska*. Accessed: Dec. 7, 2025. [Online]. Available: https://beef.unl.edu/faq-2009breedingage/.

[6] R. L. Larson and R. F. Randle, "The Bovine Estrous Cycle and Synchronization of Estrus," [Online]. Available: https://www.vet.k-state.edu/academics/student-faculty-handbook/studentorgs/aadpDocs/Estrous_Cycle_physiology1.pdf.

[7] *Physiology of Estrous Cycle in Cows & Buffaloes*, 2019. Accessed: Oct. 6, 2025. [Online]. Available: https://www.aliveterinarywisdom.com/physiology-of-estrous-cycle-in-cows-buffaloes/.

[8] I. Merkelytė, A. Šiukščius, and R. Nainienė, "The Role of Sensor Technologies in Estrus Detection in Beef Cattle: A Review of Current Applications," *Animals (Basel)*, vol. 15, no. 15, p. 2313, 2025, ISSN: 2076-2615. DOI: 10.3390/ani15152313. Accessed: Oct. 20, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12345459/.

[9] *The cost of a missed heat*. Accessed: Oct. 7, 2025. [Online]. Available: https://teagasc.ie/news--events/daily/the-cost-of-a-missed-heat/.

[10] H. A. Syah, A. P. A. Yekti, P. Utami, N. Isnaini, and T. Susilawati, "Effect of Artificial Insemination Timing on Conception Rate in Lactating Holstein-Friesian Cows," *WVJ*, vol. 14, no. 4, pp. 529–535, 2024, ISSN: 23224568. DOI: 10.54203/scil.2024.wvj60. Accessed: Oct. 12, 2025. [Online]. Available: http://wvj.science-line.com/attachments/article/83/WVJ14(4)%20529-535,%202024.pdf.

[11] *» Transition Period Management.* Accessed: Dec. 7, 2025. [Online]. Available: http://cahl.ie/transition-period-management/.

[12] F. M. Tangorra, E. Buoio, A. Calcante, A. Bassi, and A. Costa, "Internet of Things (IoT): Sensors Application in Dairy Cattle Farming," *Animals*, vol. 14, no. 21, p. 3071, 2024, ISSN: 2076-2615. DOI: 10.3390/ani14213071. Accessed: Oct. 7, 2025. [Online]. Available: https://www.mdpi.com/2076-2615/14/21/3071.

[13] "Cow comfort, behavior and welfare with specific reference to dairy cattle: A review," *Ger. J. Vet. Res.*, vol. 4, no. 3, pp. 160–175, 2024, ISSN: 2703-1322. DOI: 10.51585/gjvr.2024.3.0107. Accessed: Oct. 21, 2025. [Online]. Available: https://gmpc-akademie.de/articles/gjvr/single/222.

[14] *Classification of Cattle Behaviour and Detection of Heat(Estrus) using Sensor Data Centre for Development of Advanced Computing*. Accessed: Oct. 7, 2025. [Online]. Available: https://arxiv.org/html/2506.16380v1.

[15] D. Kaur and A. K. Virk, "Smart neck collar: IoT-based disease detection and health monitoring for dairy cows," *Discov Internet Things*, vol. 5, no. 1, p. 12, 2025, ISSN: 2730-7239. DOI: 10.1007/s43926-025-00109-5. Accessed: Oct. 7, 2025. [Online]. Available: https://doi.org/10.1007/s43926-025-00109-5.

[16] M. M. Detamo, "A Review on Estrous Detection and Associated Challenges in Farm Animals," *Journal of Cardiology Research Reviews & Reports*, vol. 4, no. 5, pp. 1–3, 2023, ISSN: 2634-6796. DOI: 10.47363/JCRRR/2023(4)201. Accessed: Oct. 7, 2025. [Online]. Available: https://www.onlinescientificresearch.com/journals/jcrrr/articles/a-review-on-estrous-detection-and-associated-challenges-in-farm-animals.html.

[17] C. J. Rutten, A. G. J. Velthuis, W. Steeneveld, and H. Hogeveen, "*Invited review:* Sensors to support health management on dairy farms," *Journal of Dairy Science*, vol. 96, no. 4, pp. 1928–1952, 2013, ISSN: 0022-0302. DOI: 10.3168/jds.2012-6107. Accessed: Oct. 12, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022030213001409.

[18] R. N. V. J. Mohan, P. S. Rayanoothala, and R. P. Sree, "Next-gen agriculture: Integrating AI and XAI for precision crop yield predictions," *Front. Plant Sci.*, vol. 15, 2025, ISSN: 1664-462X. DOI: 10.3389/fpls.2024.1451607. Accessed: Oct. 7, 2025. [Online]. Available: https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2024.1451607/full.

[19] S. Neethirajan, "Artificial Intelligence and Sensor Technologies in Dairy Livestock Export: Charting a Digital Transformation," *Sensors*, vol. 23, no. 16, p. 7045, 2023, ISSN: 1424-8220. DOI: 10.3390/s23167045. Accessed: Oct. 21, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/23/16/7045.

[20] *Dairy cows spontaneously produce milk all year round, TRUE or FALSE? - Chaire bien-être animal*, 2021. Accessed: Oct. 20, 2025. [Online]. Available: https://chaire-bea.vetagro-sup.fr/en/dairy-cows-spontaneously-produce-milk-all-year-round-true-or-false/.

[21] R. Sterry and H. Schlesser, *Estrus detection & Estrus detection aids*. Accessed: Oct. 20, 2025. [Online]. Available: https://dairy.extension.wisc.edu/articles/estrus-detection-estrus-detection-aids/.

[22] *Managing Cow Lactation Cycles*. Accessed: Oct. 20, 2025. [Online]. Available: https://www.thecattlesite.com/articles/4248/managing-cow-lactation-cycles.

[23] H. M and N. Banuvalli, "(PDF) Heat (Estrus) Detection Techniques in Dairy Farms-A Review," *ResearchGate*, 2025, ISSN: 2277-3371. Accessed: Oct. 20, 2025. [Online]. Available: https://www.researchgate.net/publication/283156028_Heat_Estrus_Detection_Techniques_in_Dairy_Farms-A_Review.

[24] A. Poborská, M. Šoch, L. Zábranský, L. Smutný, I. Novotná, and p. Smolík, "Monitoring Lameness in Cattle Using the Vitalimeter," Accessed: Oct. 12, 2025. [Online]. Available: https://www.cabidigitallibrary.org/doi/pdf/10.5555/20173012677.

[25] *Farmtec - stájové technologie*. Accessed: Oct. 20, 2025. [Online]. Available: https://www.farmtec.cz/.

[26] "Automation of oestrus detection in dairy cows: A review | Request PDF," *ResearchGate*, 2025. DOI: 10.1016/S0301-6226(01)00323-2. Accessed: Oct. 21, 2025. [Online]. Available: https://www.researchgate.net/publication/223888220_Automation_of_oestrus_detection_in_dairy_cows_A_review.

[27] F. A. M. Tuyttens, C. F. M. Molento, and S. Benaissa, "Twelve Threats of Precision Livestock Farming (PLF) for Animal Welfare," *Front. Vet. Sci.*, vol. 9, 2022, ISSN: 2297-1769. DOI: 10.3389/fvets.2022.889623. Accessed: Oct. 20, 2025. [Online]. Available: https://www.frontiersin.org/journals/veterinary-science/articles/10.3389/fvets.2022.889623/full.

[28] *Estrus (Heat) Detection in Cattle | Mississippi State University Extension Service*. Accessed: Oct. 20, 2025. [Online]. Available: https://extension.msstate.edu/publications/estrus-heat-detection-cattle.

[29] S. Reith and S. Hoy, "Review: Behavioral signs of estrus and the potential of fully automated systems for detection of estrus in dairy cattle," *Animal*, vol. 12, no. 2, pp. 398–407, 2018, ISSN: 1751-7311. DOI: 10.1017/S1751731117001975. Accessed: Oct. 20, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1751731117001975.

[30] P. L. Senger, "The estrus detection problem: New concepts, technologies, and possibilities," *J Dairy Sci*, vol. 77, no. 9, pp. 2745–2753, 1994, ISSN: 0022-0302. DOI: 10.3168/jds.S0022-0302(94)77217-9.

[31] J. B. Roelofs, F. J. C. M. van Eerdenburg, N. M. Soede, and B. Kemp, "Various behavioral signs of estrous and their relationship with time of ovulation in dairy cattle," *Theriogenology*, vol. 63, no. 5, pp. 1366–1377, 2005, ISSN: 0093-691X. DOI: 10.1016/j.theriogenology.2004.07.009. Accessed: Nov. 10, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0093691X04002456.

[32] J. B. Roelofs, F. J. C. M. Van Eerdenburg, W. Hazeleger, N. M. Soede, and B. Kemp, "Relationship between progesterone concentrations in milk and blood and time of ovulation in dairy cattle," *Animal Reproduction Science*, vol. 91, no. 3, pp. 337–343, 2006, ISSN: 0378-4320. DOI: 10.1016/j.anireprosci.2005.04.015. Accessed: Nov. 11, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378432005001077.

[33] J. E. Bagley, M. P. Richter, and T. J. Lane, "The Role of Transrectal Sonography in Pregnancy Diagnosis in Cattle," *Journal of Diagnostic Medical Sonography*, vol. 39, no. 1, pp. 50–60, 2023, ISSN: 8756-4793, 1552-5430. DOI: 10.1177/87564793221120260. Accessed: Nov. 11, 2025. [Online]. Available: https://journals.sagepub.com/doi/10.1177/87564793221120260.

[34] P. Løvendahl and M. G. G. Chagunda, "On the use of physical activity monitoring for estrus detection in dairy cows," *Journal of Dairy Science*, vol. 93, no. 1, pp. 249–259, 2010, ISSN: 0022-0302. DOI: 10.3168/jds.2008-1721. Accessed: Nov. 12, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022030210702848.

[35] P. Løvendahl and M. G. G. Chagunda, "Assessment of fertility in dairy cows based on electronic monitoring of their physical activity.," Accessed: Dec. 11, 2025. [Online]. Available: https://www.cabidigitallibrary.org/doi/full/10.5555/20063169958.

[36] E. E. At-Taras and S. L. Spahr, "Detection and Characterization of Estrus in Dairy Cattle with an Electronic Heatmount Detector and an Electronic Activity Tag1," *Journal of Dairy Science*, vol. 84, no. 4, pp. 792–798, 2001, ISSN: 0022-0302. DOI: 10.3168/jds.S0022-0302(01)74535-3. Accessed: Nov. 12, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022030201745353.

[37] T. L. Bova et al., "Environmental stressors influencing hormones and systems physiology in cattle," *Reprod Biol Endocrinol*, vol. 12, p. 58, 2014, ISSN: 1477-7827. DOI: 10.1186/1477-7827-12-58. Accessed: Nov. 12, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4094414/.

[38] M. Mičiaková, P. Strapák, E. Strapáková, and I. Szencziová, "Evaluating Rumination Time Changes During Estrus in Dairy Cows," *Dairy*, vol. 6, no. 1, p. 5, 2025, ISSN: 2624-862X. DOI: 10.3390/dairy6010005. Accessed: Nov. 12, 2025. [Online]. Available: https://www.mdpi.com/2624-862X/6/1/5.

[39] K. Hendriksen, W. Büscher, S. Hoppe, and C. Hoffmanns, "Validation of an acoustic rumination sensor for dairy cows,"

[40] G. S. Heckman, L. S. Katz, R. H. Foote, E. A. Oltenacu, N. R. Scott, and R. A. Marshall, "Estrous cycle patterns in cattle monitored by electrical resistance and milk progesterone," *J Dairy Sci*, vol. 62, no. 1, pp. 64–68, 1979, ISSN: 0022-0302. DOI: 10.3168/jds.S0022-0302(79)83203-8.

[41] C. J. Rutten, C. Kamphuis, H. Hogeveen, K. Huijps, M. Nielen, and W. Steeneveld, "Sensor data on cow activity, rumination, and ear temperature improve prediction of the start of calving in dairy cows," *Computers and Electronics in Agriculture*, vol. 132, pp. 108–118, 2017, ISSN: 0168-1699. DOI: 10.1016/j.compag.2016.11.009. Accessed: Nov. 12, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169916310262.

[42] J. Wang, M. Bell, X. Liu, and G. Liu, "Machine-Learning Techniques Can Enhance Dairy Cow Estrus Detection Using Location and Acceleration Data," *Animals*, vol. 10, no. 7, p. 1160, 2020, ISSN: 2076-2615. DOI: 10.3390/ani10071160. Accessed: Nov. 13, 2025. [Online]. Available: https://www.mdpi.com/2076-2615/10/7/1160.

[43] S. HIGAKI, H. DARHAN, C. SUZUKI, T. SUDA, R. SAKURAI, and K. YOSHIOKA, "An attempt at estrus detection in cattle by continuous measurements of ventral tail base surface temperature with supervised machine learning," *Journal of Reproduction and Development*, vol. 67, 2020. DOI: 10.1262/jrd.2020-075.

[44] Ç. M. Sakar, M. Ergin, and Y. Altay, "Comparative analysis of machine learning algorithms for estrous detection in dairy cows using sensor-based behavioral data across seasons," *Trop Anim Health Prod*, vol. 57, no. 8, p. 479, 2025, ISSN: 1573-7438. DOI: 10.1007/s11250-025-04750-8. Accessed: Nov. 13, 2025. [Online]. Available: https://doi.org/10.1007/s11250-025-04750-8.

[45] Z. Wang, Z. Hua, Y. Wen, S. Zhang, X. Xu, and H. Song, "E-YOLO: Recognition of estrus cow based on improved YOLOv8n model," *Expert Systems with Applications*, vol. 238, p. 122 212, 2024, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.122212. Accessed: Nov. 26, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423027148.

[46] I. Mienye, T. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information*, vol. 15, p. 517, 2024. DOI: 10.3390/info15090517.

[47] Y.-R. Chen, P.-Y. Chen, and C.-K. Su, "An LSTM Neural Network for Estrus Detection in Dairy Cows," in *2024 10th International Conference on Applied System Innovation (ICASI)*, 2024, pp. 196–198. DOI: 10.1109/ICASI60819.2024.10547775. Accessed: Nov. 26, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10547775.

[48] A. S. Keceli, C. Catal, A. Kaya, and B. Tekinerdogan, "Development of a recurrent neural networks-based calving prediction model using activity and behavioral data," *Computers and Electronics in Agriculture*, vol. 170, p. 105 285, 2020, ISSN: 0168-1699. DOI: 10.1016/j.compag.2020.105285. Accessed: Nov. 26, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169919312220.

[49] D. Dhakshinamoorthy, A. Jha, S. Majumdar, D. Ghosh, R. Chakraborty, and H. Ray, *Classification of Cattle Behavior and Detection of Heat (Estrus) using Sensor Data*, Comment: 6 pages, 5 figures. Druva Dhakshinamoorthy and Avikshit Jha contributed equally as co-first authors. Work conducted during a summer internship at CDAC Kolkata by students of BITS Pilani, 2025. DOI: 10.48550/arXiv.2506.16380. Accessed: Nov. 26, 2025. [Online]. Available: http://arxiv.org/abs/2506.16380.

[50] *Variation in Physical Activity as an Indication of Estrus in Dairy Cows - ScienceDirect*. Accessed: Nov. 26, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022030277838599.

[51] E. Strapáková, "Influence of estrus on changes of locomotion activity and rumination time in cattle dams," *Acta fytotechnica et zootechnica*, DOI: 10.15414/AFZ.2021.24.MI-PRAP.127-130. Accessed: Nov. 26, 2025. [Online]. Available: https://www.academia.edu/100085144/Influence_of_estrus_on_changes_of_locomotion_activity_and_rumination_time_in_cattle_dams.

[52] *The Influence of Selected Factors on Changes in Locomotion Activity during Estrus in Dairy Cows*. Accessed: Nov. 26, 2025. [Online]. Available: https://www.mdpi.com/2076-2615/14/10/1421.

[53] G. Gao et al., "CNN-Bi-LSTM: A Complex Environment-Oriented Cattle Behavior Classification Network Based on the Fusion of CNN and Bi-LSTM," *Sensors (Basel)*, vol. 23, no. 18, p. 7714, 2023, ISSN: 1424-8220. DOI: 10.3390/s23187714. Accessed: Dec. 7, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10536225/.

[54] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[55] "(PDF) Applied Logistic Regression," *ResearchGate*, 2025, ISSN: 0040-1706. Accessed: Jan. 18, 2026. [Online]. Available: https://www.researchgate.net/publication/261659875_Applied_Logistic_Regression.

[56] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn*, vol. 20, no. 3, pp. 273–297, 1995, ISSN: 1573-0565. DOI: 10.1007/BF00994018. Accessed: Dec. 9, 2025. [Online]. Available: https://doi.org/10.1007/BF00994018.

[57] A. Asif and Rashid, "(PDF) Estrus Detection in Dairy Cows from Location and Acceleration Data using Machine Learning," *ResearchGate*, 2025. DOI: 10.30537/sjcms.v6i1.1046. Accessed: Nov. 26, 2025. [Online]. Available: https://www.researchgate.net/publication/362184628_Estrus_Detection_in_Dairy_Cows_from_Location_and_Acceleration_Data_using_Machine_Learning.

[58] N. Ma, L. Pan, S. Chen, and B. Liu, "NB-IoT Estrus Detection System of Dairy Cows Based on LSTM Networks," in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, 2020, pp. 1–5. DOI: 10.1109/PIMRC48278.2020.9217214. Accessed: Nov. 26, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/9217214.

[59] *(PDF) Unleashing the Power of Python Libraries for Machine Learning Excellence (Covers Tools and Techniques for Building Smarter Models).* Accessed: Dec. 2, 2025. [Online]. Available: https://www.researchgate.net/publication/387933576_Unleashing_the_Power_of_Python_Libraries_for_Machine_Learning_Excellence_Covers_Tools_and_Techniques_for_Building_Smarter_Models.

[60] "(PDF) Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research," *ResearchGate*, 2025. DOI: 10.5336/biostatic.2022-93961. Accessed: Dec. 8, 2025. [Online]. Available: https://www.researchgate.net/publication/369458562_Class_Weighting_Technique_to_Deal_with_Imbalanced_Class_Problem_in_Machine_Learning_Methodological_Research.

[61] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal Loss for Dense Object Detection*, 2018. DOI: 10.48550/arXiv.1708.02002. Accessed: Dec. 8, 2025. [Online]. Available: http://arxiv.org/abs/1708.02002.

[62] J. Singh et al., "Batch-balanced focal loss: A hybrid solution to class imbalance in deep learning," *J Med Imaging (Bellingham)*, vol. 10, no. 5, p. 051 809, 2023, ISSN: 2329-4302. DOI: 10.1117/1.JMI.10.5.051809. Accessed: Dec. 8, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10289178/.

[63] Y. Gal and Z. Ghahramani, *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, Comment: 12 pages, 6 figures; fixed a mistake with standard error and added a new table with updated results (marked "Update [October 2016]"); Published in ICML 2016, 2016. DOI: 10.48550/arXiv.1506.02142. Accessed: Dec. 9, 2025. [Online]. Available: http://arxiv.org/abs/1506.02142.

[64] "(PDF) The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution," *ResearchGate*, 2025. DOI: 10.20982/tqmp.04.1.p013. Accessed: Jan. 4, 2026. [Online]. Available: https://www.researchgate.net/publication/49619432_The_Mann-Whitney_U_A_Test_for_Assessing_Whether_Two_Independent_Samples_Come_from_the_Same_Distribution.

[65] *Spearmanr — SciPy v1.16.2 Manual.* Accessed: Jan. 4, 2026. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html.

[66] *P-Value: What It Is, How to Calculate It, and Examples.* Accessed: Jan. 4, 2026. [Online]. Available: https://www.investopedia.com/terms/p/p-value.asp.