|  |  |
| --- | --- |
| Advancing Precision Agriculture: Machine Learning-based Early Detection of Potato Sprouting via Electrophysiological Signals | |
|  | |
| Student | Main supervisor |
| Stefano Billeter | Dr. Omran Ayoub |
| Co-supervisors |
| Dr. Davide Andreoletti |
| Company (if any) |
| - |
| Degree course | Project identifier (if any) |
| Data Science and Artificial Intelligence | DSAI\_4 |
| Year |  |
| - |  |

|  |  |
| --- | --- |
| Date |  |

|  |  |
| --- | --- |
|  |  |

Dedication

*I am profoundly grateful to my family, whose unconditional love and support have been my foundation. To my parents, thank you for believing in me and for your constant encouragement through both the good and challenging times. Your sacrifices and unwavering faith in my abilities have been my greatest motivation.*

*To my friends, thank you for being my pillars of strength. Your companionship, understanding, and words of encouragement have been a source of comfort and inspiration.*

*I would like to express my deepest gratitude to my supervisors, Dr. Omran Ayoub and Dr. Davide Andreoletti, for their guidance and continuous support throughout this project. Your expertise, encouragement and friendship have been invaluable in bringing this thesis to fruition.*

*I also wish to extend my sincere thanks to Prof. Emanuele Carpanzano, Prof. Andreas Graf, and Prof. Francesco Flammini for giving me the chance of a lifetime. I will always remember those who helped me when I needed it most.*

*This accomplishment would not have been possible without the support and belief of all of you.*

*I dedicate this work to you with heartfelt thanks.*

General index

* **Abstract**
* **Introduction (motivation to the problem, problem i want to solve, thesis objectives at high level) 3-4pgs**
* **Background 3-4pgs**
  + **On the problem**
    - **State of the art potato storage**
    - **Previous state of the art potato sprouting prevention**
    - **Actual state of the art potato sprouting prevention**
  + **On the methodology** 
    - **Methods for signal processing(FFT, wavelets, …) 2pgs each methods**
    - **Feat extraction**
    - **ML models used**
    - **UQ**
    - **XAI (shap)**
* **Problem statement and research questions 3-4pgs (task)**
* The dataset (how it is composed, methodologies to collect mV and target, devices, and so on) 3-4pgs
* Methodology 1pg overall, 10pgs per il resto
  + Preprocessing
    - Dataset cleaning, treatment of TS gaps, interpolation, windowing, …
  + Signal processing and feature extraction/generation
    - Which features, methodologies
  + Modelling
    - Which ML models, why
    - UQ
  + XAI (shap)
* Experimental settings and results, 10pgs
  + Discussion
* Conclusions
* Future steps (batch), solo per presentazione
* References

Figure index (optional, in case of many figures)

Table index ( optional, in case of many tables)

Analytical index (optional, by keywords )

**Abstract**

This project aims to develop a predictive system for detecting sprouting in potato tubers using electrophysiological signals and machine learning techniques. Given the European and Swiss ban on Chlorpropham (CIPC), a widely used anti-sprouting agent, predicting sprouting is crucial to prevent economic and quality losses during potato storage. Conducted in cooperation with Vivent SA and supported by the Swiss Confederation's Innosuisse/Agroscope, this research leverages real-time electrophysiological data from 32 potato plants.

The study utilized eXtreme Gradient Boosting (XGB), LightGBM (LGBM), and AdaBoost to predict sprouting time, focusing on feature extraction methods and window lengths. Both wavelet-based and non-wavelet-based features were tested, yielding comparable results. To enhance reliability, only pre-sprouting data was used for model training, ensuring accurate and actionable predictions for storage managers.

Future work will optimize the models, explore additional machine learning models, and develop advanced metrics for practical predictions. This research demonstrates the potential of using electrophysiological signals and machine learning for early sprouting detection, offering valuable insights for optimizing anti-sprouting treatments and minimizing storage losses.

**Introduction**

Potatoes are one of the world's most important staple crops, providing a crucial food source for millions of people globally. As a versatile and nutritious crop, potatoes are integral to diets in many countries and serve as a key agricultural product. Ensuring the quality and market value of potatoes during storage is essential for both economic stability and nutritional security. A significant challenge in potato storage is the prevention of sprouting, which can lead to substantial losses in weight, increased sugar accumulation, and reduced overall tuber quality. These factors not only affect the marketability of the potatoes but also their usability in various culinary applications.

Historically, potatoes have played a vital role in human nutrition and food security, especially during times of scarcity. Introduced to Europe from the Americas in the late 16th century, potatoes quickly became a staple food due to their high yield and nutritional value. They were crucial during periods of famine, such as the Great Famine in Ireland, and during wartime, providing a reliable source of sustenance when other crops failed. Potatoes' resilience and adaptability to various climates made them indispensable in ensuring food availability and preventing starvation.

In Swiss culture, potatoes hold a significant place in both traditional cuisine and agricultural practices. They are a fundamental ingredient in many beloved Swiss dishes, such as Rösti, Raclette, and Fondue accompaniments. The Swiss agricultural sector has long relied on potato cultivation, contributing to the nation's food self-sufficiency and rural economy. The importance of potatoes in Swiss heritage underscores the necessity of effective storage methods to maintain their quality and availability throughout the year.

**Problem Statement**

The absence of CIPC has left a significant gap in effective sprout control, resulting in increased economic losses and quality degradation during storage. Traditional methods of sprout suppression, such as using alternative chemicals or natural extracts, have proven less effective and more costly. These methods often involve frequent applications and still do not provide the level of control that CIPC once did. Consequently, there is a pressing need for innovative approaches that can accurately predict and manage sprouting events in stored potatoes, ensuring that storage conditions can be optimized to maintain the highest possible quality.

Sprouting in potatoes leads to several undesirable effects, including weight loss due to moisture loss, increased sugar content, which affects taste and cooking properties, and the overall deterioration of tuber quality. These changes not only reduce the market value of the potatoes but also pose significant challenges for processors and consumers who rely on a high-quality product. Developing a system that can provide early warnings of sprouting is crucial for minimizing these losses and ensuring that potatoes can be stored more effectively.

**Research Objectives**

This research aims to develop a predictive system for detecting sprouting in potato tubers using electrophysiological signals and advanced machine learning techniques. The specific objectives of the research are as follows:

* **Developing the Best Preprocessing Pipeline:** Optimize a comprehensive preprocessing pipeline to handle the raw electrophysiological data collected by Vivent SA. This involves preparing the data to ensure it is suitable for analysis and feature extraction.
* **Comparing Machine Learning Models:** Evaluate the efficacy of different machine learning models, including eXtreme Gradient Boosting (XGB), LightGBM (LGBM), and AdaBoost, in predicting the time-to-sprouting. This comparison will help determine the most effective model for accurate predictions.
* **Evaluating Feature Extraction Methods:** Experiment with training the models using features extracted from electrophysiological signals both with and without wavelet preprocessing. This evaluation will identify the best techniques for capturing essential characteristics of the signals and improving model performance.
* **Ensuring Model Reliability:** Focus on using only pre-sprouting data for training and validation to enhance the reliability of the predictive models. By excluding post-sprouting data, the models can avoid potential noise and confusion, leading to more accurate predictions.
* **Providing Actionable Insights:** Develop a system that offers valuable insights for storage managers to optimize the timing of anti-sprouting treatments and minimize economic losses. These insights will enable more effective management of storage conditions and treatment applications.

**Research Significance**

This project is a collaborative effort with Vivent SA, a company specializing in collecting and analyzing plant biosignals. The research is part of an Innosuisse/Agroscope initiative supported by the Swiss Confederation, emphasizing the importance of interdisciplinary collaboration and the application of advanced technologies to address real-world agricultural challenges.

The successful development of a reliable predictive system has the potential to revolutionize potato storage management. By providing early warnings of sprouting events, the system will enable storage managers to take timely and informed actions, such as applying anti-sprouting treatments precisely when needed. This proactive approach will help maintain the quality and market value of stored potatoes, reducing economic losses and ensuring a steady supply of high-quality produce.

Moreover, this research demonstrates the value of integrating electrophysiological signals with machine learning techniques to solve complex agricultural problems. The insights gained from this study can be extended to other crops and storage scenarios, showcasing the broader applicability of the methodologies developed. By addressing a critical issue in potato storage, this project contributes to the advancement of agricultural practices and the optimization of food supply chains.

**Background**

**On the problem**

Potatoes are one of the most important staple crops globally, integral to diets and economies across various regions. Their versatility, nutritional value, and adaptability to different climates have made them a crucial agricultural product. However, the challenge of storing potatoes without quality degradation, particularly preventing sprouting, remains a significant issue. Sprouting not only reduces the tuber's weight and quality but also increases sugar content, which negatively impacts their taste and cooking properties.

**State of the Art Potato Storage**

Potato storage has evolved significantly over the years. Traditional methods involved storing potatoes in cool, dark, and well-ventilated areas to prolong their shelf life and reduce sprouting. These methods were effective to some extent but did not provide complete control over the sprouting process. As technology advanced, more sophisticated storage solutions were developed.

Modern potato storage facilities now use controlled atmosphere storage (CAS) to maintain the ideal conditions for potatoes. CAS involves regulating temperature, humidity, and gas composition (such as oxygen and carbon dioxide levels) to slow down the metabolic processes of the potatoes, thereby reducing sprouting and decay. These facilities often use refrigeration systems to keep temperatures low, typically between 4°C and 10°C, which is optimal for potato storage.

Ventilation systems are crucial in modern storage facilities. They help maintain uniform temperature and humidity levels throughout the storage area, preventing hot spots where sprouting or decay could occur. Advanced sensors and monitoring systems are used to continuously track these conditions, ensuring they remain within the desired range.

While these methods have significantly improved the ability to store potatoes for extended periods, they are not foolproof. Temperature fluctuations, inadequate ventilation, and other factors can still lead to sprouting. Hence, there remains a need for more precise and reliable methods to predict and manage sprouting in stored potatoes.

**Previous State of the Art Potato Sprouting Prevention**

For many years, Chlorpropham (CIPC) was the industry standard for sprout suppression in potatoes. CIPC is a chemical compound applied to potatoes either as a fog or as a liquid treatment to inhibit the growth of sprouts. Its effectiveness and affordability made it the preferred choice for farmers and storage managers worldwide. CIPC could extend the storage life of potatoes significantly, allowing them to be stored for up to a year without significant sprouting.

However, CIPC's safety has been under scrutiny for some time. Studies indicated potential health risks associated with its residue on potatoes, leading to concerns about its long-term use. These health concerns ultimately led to regulatory actions in Europe and Switzerland, resulting in the ban of CIPC in these regions. The ban created a significant gap in sprouting prevention methods, as no other chemical treatments were as effective or affordable as CIPC.

Before the ban, other methods such as mechanical ventilation and temperature control were often used in conjunction with CIPC to maximize sprout suppression. These methods alone, however, were not as effective in controlling sprouting, leading to increased interest in finding new solutions post-CIPC.

**Actual State of the Art Potato Sprouting Prevention**

In the wake of the CIPC ban, the potato industry has been exploring alternative sprouting prevention methods. Several new approaches and technologies have been developed, though each comes with its own set of challenges and limitations.

Chemical Alternatives: With CIPC banned, alternative chemicals like ethylene, spearmint oil, and orange oil have been introduced for sprout suppression. Ethylene, a naturally occurring plant hormone, can be used to inhibit sprout growth when applied in controlled doses. However, its effectiveness varies depending on the potato variety and storage conditions. Similarly, natural oils such as spearmint and orange oil have been found to suppress sprouting, but they are generally less effective and more expensive than CIPC.

Physical Methods: These include the use of ultraviolet (UV) light and thermal treatments. UV light can inhibit sprout growth by damaging the sprout tissues, while thermal treatments involve exposing potatoes to brief periods of high temperatures to kill developing sprouts. Both methods are still in experimental stages and have not been widely adopted due to inconsistent results and potential damage to the potatoes.

Biological Control: This involves using natural predators or pathogens to control sprout growth. Research is ongoing into the use of bacteria and fungi that can inhibit sprouting without harming the potatoes. While promising, these methods are not yet commercially viable and require further development.

Electrophysiological Monitoring and Machine Learning: The most cutting-edge approach involves using electrophysiological signals from potato plants to monitor and predict sprouting. Electrophysiological signals provide real-time data on the physiological state of the potatoes, which can be analyzed using advanced machine learning algorithms to predict sprouting events before they become visible. This method offers a non-invasive, accurate, and timely solution to sprouting prevention. By detecting changes in the potatoes' electrical signals, this approach can provide early warnings and allow for timely interventions, such as adjusting storage conditions or applying sprout suppressants.

This innovative method leverages technology to provide more precise and actionable insights into the sprouting process. Companies like Vivent SA are at the forefront of this research, working on integrating electrophysiological monitoring with advanced data analysis techniques. This approach not only addresses the immediate need for effective sprout suppression post-CIPC but also opens new avenues for improving potato storage management through technology.

While traditional methods and new chemical treatments provide some level of control over sprouting, they are not without limitations. The development of electrophysiological monitoring combined with machine learning represents a significant advancement in the field, offering a more reliable and sophisticated solution to the problem of potato sprouting during storage. This approach underscores the potential of interdisciplinary collaboration and technological innovation in solving agricultural challenges, paving the way for more sustainable and efficient storage practices.

**Background**

**On the methodology**

The methodology for this research involves several key components, including signal processing, feature extraction, machine learning models, uncertainty quantification, and explainable AI techniques. Each of these components plays a crucial role in developing a predictive system for detecting sprouting in potato tubers.

**Methods for Signal Processing**

Signal processing is a critical step in preparing electrophysiological data for analysis. The two primary methods used in this study are the Fast Fourier Transform (FFT) and Continuous Wavelet Transform (CWT).

**Fast Fourier Transform (FFT):**

FFT is used to convert the time-domain signals into the frequency domain, allowing the analysis of the signal's frequency components. This transformation is essential for identifying patterns that may not be apparent in the time domain. In this study, FFT is specifically used to extract features such as total energy, power ratio between low and high frequencies, and peak frequencies. The FFT helps in capturing the spectral properties of the signal, which are crucial for understanding the overall behavior of the potatoes' physiological state.

FFT operates by breaking down a signal into its constituent sinusoids of different frequencies. By doing so, it reveals the frequency spectrum hidden within the time-domain data. This is particularly useful for detecting periodic patterns and identifying the dominant frequencies present in the electrophysiological signals. In practical terms, FFT provides a way to decompose complex signals into simpler components, which can then be analyzed individually.

For example, in our study, the FFT was applied to the raw electrophysiological signals to extract features like the total energy of the signal, which is indicative of the overall activity within the potato tuber. The power ratio between low and high frequencies provides insight into the balance of different types of activity, while peak frequencies highlight the most prominent oscillatory behaviors.

**Continuous Wavelet Transform (CWT):**

CWT is applied to analyze localized variations of power within the signal. Unlike FFT, which provides global frequency information, CWT offers time-frequency localization, making it suitable for detecting transient features in the signal. In this study, the wavelet function used is the Morlet wavelet, chosen for its effectiveness in representing non-stationary signals. The entire electrophysiological signal is decomposed using wavelets, and this decomposition is compared with analyses performed without wavelets. The scale of the wavelet is determined based on the desired frequency range and the sampling period. This approach allows for a detailed examination of the signal's temporal and spectral characteristics.

Wavelet transforms differ from Fourier transforms in that they provide a time-frequency representation of the signal. This means that wavelets can analyze both the frequency content and its evolution over time. This dual capability makes CWT particularly powerful for analyzing non-stationary signals where the frequency components change over time, as is often the case with biological signals.

In our study, the Morlet wavelet was used due to its Gaussian shape and good time-frequency localization properties. The CWT decomposes the signal into different frequency components at various scales, allowing for the identification of transient features that may be associated with the onset of sprouting. By comparing wavelet-transformed signals with and without certain features, we can determine which aspects of the signal are most predictive of sprouting events.

**Feature Extraction**

Feature extraction is performed to convert the processed signals into meaningful metrics that can be used as inputs for machine learning models. The features extracted from the signals include a combination of statistical, frequency, and time-domain characteristics.

**Statistical Features:**

These include basic statistical measures such as mean, standard deviation, variance, minimum, maximum, range, median, skewness, and kurtosis. These features provide a summary of the signal's overall behavior.

Mean: Represents the average value of the signal over time, indicating the central tendency.

Standard Deviation and Variance: Measure the dispersion or spread of the signal values, reflecting the variability.

Minimum and Maximum: Indicate the extreme values within the signal, showing the range of activity.

Range: The difference between the maximum and minimum values, providing a simple measure of variability.

Median: The middle value of the signal, offering a robust measure of central tendency that is less affected by outliers.

Skewness: Measures the asymmetry of the signal distribution, indicating whether the signal is biased towards higher or lower values.

Kurtosis: Measures the peakedness of the signal distribution, highlighting the presence of extreme values or outliers.

Statistical features are essential as they provide a broad understanding of the signal’s characteristics, which can be crucial for differentiating between sprouting and non-sprouting states.

**Frequency Features:**

Derived from the FFT, these include total energy, power ratio between low and high frequencies, and peak frequencies. These features help capture the signal's spectral properties.

Total Energy: The sum of the squared amplitudes of the frequency components, representing the overall power of the signal.

Power Ratio (Low/High): The ratio of the power in low-frequency components to the power in high-frequency components, indicating the balance of different types of activity.

Peak Frequencies: The frequencies with the highest amplitudes, identifying the dominant oscillatory behaviors within the signal.

Frequency features provide insights into the rhythmic and periodic elements of the signal, which can be directly related to physiological processes within the potato tubers.

**Time-Domain Features:**

These include zero-crossing rate, mean-crossing rate, and local extrema rates (maxima and minima). These features describe the signal's oscillatory behavior.

Zero-Crossing Rate: The rate at which the signal crosses the zero axis, indicating the frequency of oscillations.

Mean-Crossing Rate: Similar to the zero-crossing rate but uses the mean value of the signal as the reference point.

Local Maxima and Minima Rates: The rates at which the signal reaches local high and low points, reflecting the frequency of peaks and troughs in the signal.

Time-domain features are crucial for understanding the dynamics of the signal over time, providing context to the variations observed in the signal.

Quantile Features:

Quantiles (25th, 50th, and 75th percentiles) and interquartile ranges are calculated to understand the distribution of the signal values.

Quantiles: Provide a way to divide the signal values into intervals with equal probabilities, offering a detailed view of the signal distribution.

Interquartile Range (IQR): The range between the 25th and 75th percentiles, indicating the spread of the middle 50% of the signal values.

Quantile features help in understanding the statistical distribution of the signal values, providing insights into the range and spread of the data.

Envelope Features:

Using the Hilbert transform, the signal envelope is calculated to capture its amplitude variations. Features such as envelope mean, standard deviation, maximum, minimum, skewness, and kurtosis are extracted.

Hilbert Transform: Used to compute the analytic signal, from which the envelope is derived.

Envelope Mean, Std, Max, Min, Skewness, Kurtosis: These features provide a summary of the amplitude variations within the signal, highlighting patterns that may be related to physiological changes.

Envelope features are particularly useful for capturing the intensity and variability of the signal, providing additional context to the oscillatory and frequency-based features.

The extracted features are used to train machine learning models, providing the necessary input for accurate predictions.

**Machine Learning Models Used**

Several machine learning models are employed to predict the time-to-sprouting based on the extracted features. The models used in this study include:

eXtreme Gradient Boosting (XGB):

A powerful ensemble learning method known for its high performance in predictive modeling. It builds multiple decision trees sequentially, each one correcting the errors of its predecessor. XGB is particularly effective in handling complex relationships and interactions between features, making it suitable for this application.

XGB operates by creating an ensemble of weak learners (decision trees) that are combined to form a strong predictor. Each tree is built to minimize the errors of the previous trees, leading to a highly accurate and robust model. The flexibility of XGB allows it to handle various types of data and capture intricate patterns, which is crucial for predicting sprouting events based on electrophysiological signals.

LightGBM (LGBM):

Another gradient boosting framework that is efficient and scalable. LGBM is designed to handle large datasets with high speed and performance. It uses a histogram-based algorithm to bin continuous features into discrete bins, reducing memory usage and computation time.

LGBM builds decision trees in a leaf-wise manner, which allows it to grow more complex trees with fewer splits. This results in a more accurate model that can capture intricate patterns in the data. The efficiency of LGBM makes it suitable for real-time applications where quick predictions are needed.

AdaBoost:

An ensemble method that combines the outputs of weak classifiers to create a strong classifier. It adjusts the weights of misclassified instances, focusing on difficult cases in subsequent iterations. AdaBoost is particularly useful for improving the performance of simple models by focusing on the most challenging instances.

AdaBoost operates by training multiple weak classifiers (e.g., decision stumps) and combining their predictions through a weighted majority vote. Each classifier is trained to correct the mistakes of the previous ones, leading to a strong overall model. The adaptability of AdaBoost makes it suitable for handling diverse and complex datasets.

These models are configured with standard hyperparameters, and their performance is evaluated based on metrics such as Mean Absolute Error (MAE) and Average Delta Days (ADD).

Metrics for Model Evaluation

In this study, the performance of the machine learning models is evaluated using two key metrics: Mean Absolute Error (MAE) and Average Delta Days (ADD). While MAE is a widely recognized metric, ADD is a custom metric designed specifically for this research.

Mean Absolute Error (MAE):

MAE is a measure of the average magnitude of errors between the predicted and actual values. It is calculated as the average of the absolute differences between predicted and actual values.

Immagine che contiene Carattere, bianco, diagramma, schizzo

Descrizione generata automaticamente

MAE provides a straightforward interpretation of the prediction accuracy, with lower values indicating better performance. It is particularly useful for understanding the average error in the predictions, making it an ideal metric for regression tasks like this one.

Average Delta Days (ADD):

ADD is a custom metric developed for this study to measure the average deviation in days between the predicted sprouting dates and the actual sprouting dates. This metric is crucial for practical applications where the timing of predictions is as important as the accuracy of the predicted values.

The implementation of ADD involves the following steps:

Prediction of Dates: For each test instance, the predicted time-to-sprouting is added to the initial date to estimate the sprouting date.

Calculation of Delta Days: The difference in days between the predicted sprouting date and the actual sprouting date is calculated.

Averaging the Deltas: The average of these differences (delta days) is computed across all instances to provide a single measure of timing accuracy.

ADD is defined as:

Immagine che contiene Carattere, bianco, testo, diagramma

Descrizione generata automaticamente

This metric provides a clear indication of how close the model's predictions are to the actual sprouting events in terms of days, which is critical for making timely and effective decisions in storage management.

Uncertainty Quantification (UQ)

Uncertainty quantification is integrated into the machine learning models to assess the reliability of the predictions. This involves estimating the uncertainty associated with each prediction, which helps in making informed decisions.

Techniques Used:

The primary technique used for quantifying uncertainty in this study is bootstrapping. Multiple models are trained using resampled datasets to capture the variability in the predictions. Specifically, 10 bootstrap models are trained, and their predictions are aggregated to estimate the mean and variance of the predictions. This method provides a measure of confidence in the predictions by analyzing the variability in the model outputs.

Bootstrapping involves creating multiple resampled versions of the training data by randomly sampling with replacement. Each resampled dataset is used to train a separate model, resulting in a set of models that capture the variability in the data. By aggregating the predictions of these models, we can estimate the uncertainty of the predictions.

Integration into Models:

The uncertainty estimates are incorporated into the decision-making process, allowing the models to flag predictions with high uncertainty. The predictions are categorized into certain and uncertain based on a predefined variance threshold. Predictions with variance below the threshold are considered reliable, while those above the threshold are flagged as uncertain. This ensures that only reliable predictions are used for further actions, enhancing the robustness of the model.

**The implementation involves the following steps:**

Train Bootstrap Models: Train multiple models using resampled datasets to capture prediction variability. This step ensures that the model captures the inherent uncertainty in the data.

Predict with Uncertainty: Aggregate the predictions from the bootstrap models to estimate mean and variance. Identify certain and uncertain predictions based on the variance threshold. This step provides a measure of confidence in the predictions, allowing for more informed decision-making.

**Explainable AI (XAI) - SHAP**

Explainable AI techniques, specifically SHapley Additive exPlanations (SHAP), are used to interpret the predictions of the machine learning models.

SHAP Values:

SHAP values provide a measure of the contribution of each feature to the prediction. By decomposing the prediction into additive contributions from each feature, SHAP helps in understanding the model's decision-making process. This is crucial for validating the model and ensuring its transparency.

SHAP values are based on cooperative game theory and provide a way to fairly distribute the prediction among the features. Each feature's contribution is calculated by considering all possible combinations of features, ensuring that the importance of each feature is accurately represented.

Insights Gained:

SHAP values are used to identify the most influential features, providing insights into the factors driving sprouting in potato tubers. For example, features such as mean, standard deviation, and specific frequency components might be highlighted as key predictors. This interpretability allows for a better understanding of the underlying physiological processes and can inform improvements in the model.

By analyzing SHAP values, we can gain insights into the relationships between features and the target variable. This information can be used to refine the feature extraction process, improve model performance, and ensure that the model's predictions are based on meaningful and interpretable patterns.

The combination of these methodologies forms a robust framework for predicting potato sprouting, leveraging advanced signal processing, feature extraction, machine learning, uncertainty quantification, and explainable AI techniques to deliver accurate and actionable insights.

**Problem statement and research questions**

The storage of potatoes, one of the world's most vital staple crops, poses significant challenges, particularly in preventing sprouting during extended storage periods. Historically, the agricultural industry has relied heavily on chemical agents like Chlorpropham (CIPC) to inhibit sprouting. CIPC's effectiveness and affordability made it the standard for sprout suppression in potatoes, ensuring their quality and marketability during storage. However, recent health and environmental concerns have led to regulatory bans on CIPC in Europe and Switzerland. This development has created an urgent need for alternative sprout suppression methods that are both effective and safe.

The absence of CIPC has left a significant gap in the tools available for controlling sprout growth. Traditional methods of sprout suppression, which often involve alternative chemical treatments, have proven less effective and more costly. These methods typically require frequent applications and still do not provide the level of control that CIPC once did. Consequently, there is a pressing need for innovative approaches that can accurately predict and manage sprouting events in stored potatoes, ensuring that storage conditions can be optimized to maintain the highest possible quality.

Sprouting in potatoes leads to several undesirable effects, including weight loss due to moisture evaporation, increased sugar content that affects taste and cooking properties, and overall deterioration of tuber quality. These changes not only reduce the market value of the potatoes but also pose significant challenges for processors and consumers who rely on high-quality produce. Therefore, developing a system that can provide early warnings of sprouting is crucial for minimizing these losses and ensuring that potatoes can be stored more effectively.

**Research Questions**

To address the problem of potato sprouting during storage in the absence of CIPC, this research aims to develop a predictive system using electrophysiological signals and advanced machine learning techniques. The following research questions guide this study:

1. **How can electrophysiological signals be effectively used to monitor and predict sprouting in potato tubers?**
   * This question focuses on the potential of using non-invasive electrophysiological monitoring to gather real-time data from potato tubers. By analyzing these signals, the study aims to identify patterns indicative of the physiological changes that precede sprouting.
2. **How can the combination of various feature extraction methods from electrophysiological signals improve the prediction accuracy of sprouting events in potato tubers?**

* This question investigates the effectiveness of integrating different feature extraction techniques, including statistical measures, Fast Fourier Transform (FFT)-derived frequency features, and wavelet-derived time-frequency features. The goal is to determine how the combined use of these methods can enhance the accuracy and reliability of predicting sprouting events.

1. **Which machine learning models are best suited for predicting the time-to-sprouting in potatoes based on the extracted features?**
   * This question evaluates the performance of various machine learning models, including eXtreme Gradient Boosting (XGB), LightGBM (LGBM), and AdaBoost, in predicting sprouting. By comparing these models, the study seeks to identify the most accurate and reliable approach for sprout prediction.
2. **How can uncertainty quantification be integrated into the predictive models to enhance their reliability and decision-making process?**
   * This question explores techniques for quantifying the uncertainty in the model predictions. By integrating uncertainty measures, the study aims to ensure that only reliable predictions are used for decision-making, thereby improving the robustness of the predictive system.
3. **What insights can be gained from explainable AI techniques, such as SHAP (SHapley Additive exPlanations), about the factors driving sprouting in potatoes?**
   * This question examines the application of explainable AI techniques to interpret the machine learning model predictions. Understanding which features contribute most to the predictions can provide valuable insights into the physiological processes underlying sprouting and help refine the predictive models.

**Detailed Discussion**

The problem of potato sprouting during storage is multifaceted, involving biological, environmental, and technological aspects. The transition away from CIPC necessitates a comprehensive understanding of these factors to develop effective alternative solutions.

Biological Factors: Potato tubers undergo complex physiological changes during storage that can lead to sprouting. These changes are influenced by factors such as temperature, humidity, light exposure, and the tubers' metabolic activity. Electrophysiological signals, which capture the electrical activity within the tubers, offer a window into these physiological processes. By monitoring these signals, it is possible to detect early signs of sprouting and intervene before visible sprouts appear.

Environmental Factors: The storage environment plays a critical role in sprout suppression. Optimal conditions for potato storage include low temperatures, high humidity, and minimal light exposure. However, maintaining these conditions consistently can be challenging. Variations in the storage environment can trigger sprouting, making it essential to develop a system that can adapt to changing conditions and provide timely alerts for necessary adjustments.

Technological Factors: Advances in sensor technology and data analysis techniques have opened new avenues for monitoring and managing potato storage. Electrophysiological sensors can continuously record data from the tubers, providing real-time insights into their physiological state. Machine learning models can analyze this data to predict sprouting events, while explainable AI techniques can elucidate the underlying factors driving these predictions.

The integration of these factors into a cohesive predictive system involves several steps:

1. **Data Collection:** Electrophysiological signals are collected from potato tubers using specialized sensors. This data forms the basis for further analysis and model development.
2. **Signal Processing:** The raw signals are processed using techniques such as FFT and CWT to extract meaningful features. FFT provides global frequency information, while CWT offers time-frequency localization, capturing transient features in the signal.
3. **Feature Extraction:** Relevant features are extracted from the processed signals, including statistical measures, frequency components, and time-domain characteristics. These features serve as inputs for the machine learning models.
4. **Model Training and Validation:** Machine learning models, including XGB, LGBM, and AdaBoost, are trained on the extracted features to predict the time-to-sprouting. The models are validated using cross-validation techniques to ensure their accuracy and robustness.
5. **Uncertainty Quantification:** Techniques such as bootstrapping are used to quantify the uncertainty in the model predictions. By aggregating the predictions from multiple models, the study assesses the reliability of the predictions and identifies certain and uncertain cases.
6. **Explainable AI:** SHAP values are computed to interpret the model predictions, highlighting the most influential features. This transparency helps validate the models and provides insights into the physiological processes driving sprouting.

By addressing these steps, the research aims to develop a reliable and actionable system for predicting potato sprouting, ensuring that storage conditions can be optimized to maintain the highest possible quality of the potatoes.

In conclusion, the problem of potato sprouting during storage in the absence of CIPC presents a significant challenge that requires innovative solutions. This research seeks to develop a predictive system based on electrophysiological signals and advanced machine learning techniques, guided by key research questions that address the critical aspects of signal processing, model development, uncertainty quantification, and explainability. The successful implementation of this system has the potential to revolutionize potato storage management, offering early warnings of sprouting events and enabling timely interventions to maintain the quality and marketability of stored potatoes.

**The dataset**

The dataset used in this study is a comprehensive collection of electrophysiological signals from 32 potato tubers, provided by Vivent SA. This dataset is crucial for developing a predictive system for detecting sprouting in potato tubers. The detailed composition and characteristics of the dataset are as follows:

**Data Collection**

Vivent SA, a company specializing in plant biosignals, collected the electrophysiological data from potato tubers. Each of the 32 potato tubers was equipped with sensors to monitor their electrophysiological activity. The data was collected with a frequency of one second, capturing real-time changes in the electrical signals within the tubers.

**Electrophysiological Signals:**

* **Sampling Frequency:** The data was sampled at a frequency of 1 Hz (one sample per second), providing a high-resolution view of the electrophysiological activity.
* **Measurement Unit:** The electrical signals were measured in millivolts (mV), reflecting the voltage changes within the potato tubers.

**Sprouting Event Annotation:**

* Each potato in the dataset includes a timestamp indicating the exact moment of sprouting. This annotation is crucial for aligning the electrophysiological signals with the sprouting events, allowing for precise analysis and model training.

**Dataset Composition**

The dataset is organized into 32 dataframes, one for each potato tuber. Each dataframe contains two primary components: the electrophysiological signals and the sprouting event annotation.

**1. Electrophysiological Signals:**

* **Timestamp:** The time at which each measurement was taken. The timestamp is crucial for temporal analysis and aligning the signals with the sprouting event.
* **Voltage Measurements:** The voltage readings (in millivolts) taken at each timestamp. These readings capture the electrophysiological activity within the potato tubers, providing the raw data needed for signal processing and feature extraction.

**2. Sprouting Event Annotation:**

* **Days to Sprouting:** The number of days from each timestamp to the sprouting event. This is a calculated field that helps in understanding the temporal proximity of each measurement to the sprouting event.

An example of the data format for each dataframe is as follows:

| **Timestamp** | **Voltage (mV)** | **Days to Sprouting** |
| --- | --- | --- |
| 2023-01-01 00:00:00 | 0.85 | 5 |
| 2023-01-01 00:00:01 | 0.86 | 5 |
| ... | ... | ... |
| 2023-01-05 12:34:56 | 1.05 | 0 |
|  |  |  |

In this example, the "Days to Sprouting" column represents the integer number of days remaining until the sprouting event from each timestamp.

**Data Characteristics**

**Temporal Resolution:**

* The high temporal resolution (1-second sampling) allows for detailed analysis of the electrophysiological signals, capturing rapid changes and transient events that may be associated with sprouting.

**Data Volume:**

* Given the high sampling frequency, the dataset for each potato tuber contains a large number of data points. For instance, a single day of recordings at 1 Hz results in 86,400 data points per potato. Over the course of several days or weeks, this accumulates to a substantial volume of data, making it rich for analysis but also challenging in terms of data handling and processing.

**Variability:**

* The electrophysiological signals exhibit variability due to natural biological differences between the potato tubers and environmental factors. This variability is essential for training robust machine learning models that can generalize well to new data.

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is performed to gain insights into the dataset and identify patterns related to sprouting. Key aspects of EDA include:

**Descriptive Statistics:**

* Summary statistics such as mean, standard deviation, and range are calculated for the voltage measurements to understand the overall distribution and variability of the signals. These statistics help in identifying the central tendency, spread, and any potential outliers in the data.

**Temporal Patterns:**

* The temporal evolution of the electrophysiological signals is analyzed to identify changes and trends leading up to the sprouting events. Time series plots are used to visualize the signal variations over time, helping to spot any periodic or sudden changes that may indicate the onset of sprouting.

**Data Challenges and Considerations**

**Variability and Noise:**

* The electrophysiological signals exhibit natural variability due to biological differences and environmental factors. Additionally, noise and artifacts may be present in the raw data, requiring careful attention during analysis.

**Temporal Alignment:**

* Ensuring accurate temporal alignment of the signals and sprouting events is crucial for reliable analysis. Any misalignment can lead to incorrect interpretations and model predictions.

**Data Volume:**

* Handling the large volume of data efficiently is a significant challenge. Each potato generates 86,400 data points per day, leading to millions of data points over the collection period. Efficient data storage, retrieval, and processing techniques are essential to manage this volume of data.

**Data Imbalance:**

* While not typically a concern in regression tasks, it is important to note that the sprouting event itself is a rare occurrence relative to the continuous stream of data collected. Ensuring the models are trained on sufficient examples of pre-sprouting activity is crucial for accurate predictions.

**Data Integrity:**

* Ensuring the integrity and quality of the data collected from the sensors is vital. Any loss or corruption of data can impact the analysis and subsequent model training. Regular checks and validations are required to maintain data integrity.

**Summary**

The dataset provided by Vivent SA forms the foundation of this research, offering a rich source of electrophysiological signals from 32 potato tubers. With high temporal resolution and precise sprouting event annotations, the dataset enables detailed analysis and model training for sprout prediction. Through careful exploration and visualization, the dataset is prepared for feature extraction and machine learning model development, paving the way for accurate and actionable predictions.

By leveraging this comprehensive dataset, the research aims to develop a robust predictive system that can effectively detect sprouting in potato tubers, addressing a critical challenge in agricultural storage management. The high-resolution data allows for the capture of subtle changes in the tubers' physiological states, which can be used to predict sprouting events accurately.

**Practical Implications**

The implications of this dataset extend beyond academic research. In practical terms, developing a predictive system based on this data can significantly impact potato storage management. By predicting sprouting events accurately, storage managers can take timely actions to prevent sprouting, such as adjusting storage conditions or applying alternative sprout inhibitors.

**Economic Impact:**

* Preventing sprouting can lead to significant cost savings by reducing spoilage and maintaining the quality of stored potatoes. This has direct economic benefits for farmers and storage facilities.

**Quality Control:**

* Ensuring the quality of potatoes during storage is crucial for maintaining their market value. Predictive systems can help in maintaining high standards of quality control, leading to better products for consumers.

**Environmental Benefits:**

* Reducing the need for chemical sprout inhibitors through accurate prediction and timely intervention can have positive environmental impacts. It reduces the reliance on chemicals and promotes more sustainable agricultural practices.

****Appendices****

They are included directly in the documentation file.

They must be identified individually with A1, A2, A3, etc. and related titles, or in groups with A1, A2, etc., B1, B2, etc., C1, C2, etc., in case you want to highlight blocks of appendices of the same type, such as schematics, diagrams, listings, etc.

In the case of appendices on which it is impossible or too complicated to insert page numbering, such as for example already formatted listings, a first numbered page must be inserted which shows the content and the number of pages of the appendix itself, then followed by the original pages (with self-numbering or without numbering).

****Attachments****

They are part of documentation , but not of the related file, since it is separate material, even if referring to the documentation itself.

This is the CD containing the documentation itself and other material related to the project, ev . separate files (such as an operating manual), ev . experimental material of the project.

The identification takes place as for the Appendices, but using different initial letters, so as not to confuse the Appendices and the Annexes.

Important: the Annexes, being separate, must always be accompanied by writings or labels that identify them as relating to the project and the documentation to which they refer (title, code, etc.).