

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**MACHINE LEARNING-BASED PREDICTION OF FTIR SPECTRAL PEAKS
FOR BIOMASS CHARACTERIZATION**



M.Sc. THESIS

Fahreddin Talha SAĞIŞ

Department of Chemical Engineering

Chemical Engineering Programme

JUNE 2025

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**MACHINE LEARNING-BASED PREDICTION OF FTIR SPECTRAL PEAKS
FOR BIOMASS CHARACTERIZATION**

M.Sc. THESIS

**Fahreddin Talha SAĞIŞ
(506221016)**

Department of Chemical Engineering

Chemical Engineering Programme

Thesis Advisor: Prof. Dr. Serdar YAMAN

JUNE 2025

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**BİYOKÜTLE KARAKTERİZASYONU İÇİN FTIR SPEKTRAL PİK
NOKTALARININ MAKİNE ÖĞRENMESİ TABANLI TAHMİNİ**

YÜKSEK LİSANS TEZİ

**Fahreddin Talha SAĞIŞ
(506221016)**

Kimya Mühendisliği Bölümü

Kimya Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Serdar YAMAN

HAZİRAN 2025

Fahreddin Talha Sağış, a M.Sc. student of İTÜ Graduate School student ID 506221016, successfully defended the thesis/dissertation entitled “MACHINE LEARNING-BASED PREDICTION OF FTIR SPECTRAL PEAKS FOR BIOMASS CHARACTERIZATION”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Prof. Dr. Serdar YAMAN**
İstanbul Technical University

Jury Members : **Prof. Dr. Serdar YAMAN**
İstanbul Technical University

.....
Prof. Dr. Hanzade AÇMA
İstanbul Technical University

.....
Assoc. Prof. Dr. Halit Eren FİGEN
Yıldız Technical University

Date of Submission : 7 May 2025
Date of Defense : 23 June 2025





To my family, friends and sufle (cat),



FOREWORD

This thesis represents the final step of my M.Sc. studies at Istanbul Technical University and reflects my efforts to integrate data-driven methods with biomass characterization.

I would like to sincerely thank my advisor, Prof. Dr. Serdar YAMAN, for his valuable guidance, encouragement, and scientific insight throughout the research process. His expertise was instrumental in shaping the direction and depth of this study.

I also extend my gratitude to my family and friends for their unwavering support during this process.

June 2025

Fahreddin Talha SAĞIŞ
(Chemical Engineer)



TABLE OF CONTENTS

	<u>Page</u>
FOREWORD.....	ix
TABLE OF CONTENTS.....	xi
ABBREVIATIONS	xiii
SYMBOLS	xv
LIST OF TABLES	xvii
LIST OF FIGURES	xix
SUMMARY	xxi
ÖZET	xxiii
1. INTRODUCTION.....	1
1.1 Background of Biomass Characterization.....	1
1.2 Importance of FTIR Analysis in Biomass Research	4
1.3 Motivation for Machine Learning Applications.....	6
1.4 Research Objectives and Hypotheses.....	8
1.5 Scope of the Study.....	10
1.6 Thesis Structure.....	12
2. LITERATURE REVIEW.....	15
2.1 Biomass Composition and Analysis Techniques	15
2.1.1 Conventional analysis methods.....	16
2.1.2 FTIR spectroscopy for biomass characterization.....	16
2.2 Machine Learning for Spectral Data	18
2.2.1 Multivariate regression models.....	19
2.2.2 Classification models	21
2.2.3 Unsupervised feature extraction.....	22
2.2.4 Data preprocessing and other techniques.....	23
2.2.5 Illustrative applications in literature	24
2.3 Research Gap and Contribution	25
2.3.1 Limitations of existing approaches	26
2.3.2 Adoption of advanced ML algorithms for improved accuracy.....	29
2.3.3 Enhanced model generality and robustness	30
2.3.4 Interpretability and spectral insight.....	30
2.3.5 Integrative improvement and practicality	31
3. MATERIALS AND METHODS	33
3.1 Biomass Sample Collection and Preparation	33
3.2 FTIR Spectroscopy.....	34
3.3 Machine Learning Approach.....	35
3.3.1 Dataset construction.....	36
3.3.2 Machine learning model selection	39
3.3.3 Feature engineering & data preprocessing.....	40
3.3.4 Training & validation.....	41
4. RESULTS AND DISCUSSION	43
4.1 Model Performance (Phase 1: Full-Spectrum Regression)	43

4.1.1 Regression metrics and comparisons	43
4.1.2 Visualizing predicted vs. actual spectra	45
4.1.3 Interpretation and discussion.....	46
4.2 Model Performance (Phase 2: Broad-Range Classification)	47
4.2.1 Overall classification metrics	48
4.2.2 Confusion matrices by interval	50
4.2.3 Effect of preprocessing and polynomial expansion	53
4.2.4 Comparison of ML models	54
4.2.5 Conclusions from Phase 2	55
4.3 Model Performance (Phase 3: Narrow-Range Classification).....	56
4.3.1 Overall classification metrics	57
4.3.2 Confusion matrices by interval	58
4.3.3 Discussion and insights	62
4.3.4 Conclusions from Phase 3	63
4.4 Discussion.....	64
5. CONCLUSIONS AND FUTURE WORK	69
5.1 Summary of Key Findings.....	69
5.2 Scientific Contributions.....	69
5.3 Limitations of the Study	70
5.4 Future Research Directions	71
REFERENCES	75
APPENDICES	79
APPENDIX A: Biomass Analysis Results.....	79
CURRICULUM VITAE	83

ABBREVIATIONS

AI	: Artificial Intelligence
ANN	: Artificial Neural Network
ATR	: Attenuated Total Reflectance
FTIR	: Fourier Transform Infrared
HCA	: Hierarchical Cluster Analysis
IR	: Infrared
KNN	: k-Nearest Neighbors
LDA	: Linear Discriminant Analysis
ML	: Machine Learning
MLP	: Multi-Layer Perceptron
NIR	: Near Infrared
NMR	: Nuclear Magnetic Resonance
NN	: Neural Network
PCA	: Principal Component Analysis
PLS	: Partial Least Squares
PLSDA	: Partial Least Squares Discriminant Analysis
PLSR	: Partial Least Squares Regression
RBF	: Radial Basis Function
RF	: Random Forest
RMSE	: Root Mean Square Error
RMSEP	: Root Mean Square Error of Prediction
RPD	: Ratio of Performance to Deviation
SVM	: Support Vector Machine
SVR	: Support Vector Regression
UV	: Ultraviolet
SHAP	: SHapley Additive exPlanations
SMOTE	: Synthetic Minority Over-Sampling Technique



SYMBOLS

°C	: Degrees Celsius
cm⁻¹	: Wavenumber (inverse centimeters)
R²	: Coefficient of determination
RMSE	: Root Mean Square Error





LIST OF TABLES

	<u>Page</u>
Table 3.1 : FTIR spectra intervals.....	37
Table A.1 : Biomass analysis results.....	80





LIST OF FIGURES

	<u>Page</u>
Figure 4.1 : Comparison of test RMSE across models.	44
Figure 4.2 : Comparison of test R2 across models.	45
Figure 4.3 : Comparison of true vs predicted FTIR spectrum for a test sample	46
Figure 4.4 : Comparison of model evaluation metrics.....	49
Figure 4.5 : Confusion matrices for gradient boosting.	50
Figure 4.6 : Confusion matrices for logistic regression.	50
Figure 4.7 : Confusion matrices for random forest.....	51
Figure 4.8 : Confusion matrices for SVM (RBF kernel).	51
Figure 4.9 : Comparison of model evaluation metrics.....	58
Figure 4.10 : Confusion matrices for gradient boosting.	58
Figure 4.11 : Confusion matrices for logistic regression.	59
Figure 4.12 : Confusion matrices for random forest.	60
Figure 4.13 : Confusion matrices for SVM (RBF kernel).	60



MACHINE LEARNING-BASED PREDICTION OF FTIR SPECTRAL PEAKS FOR BIOMASS CHARACTERIZATION

SUMMARY

This thesis explores the integration of machine learning (ML) with Fourier Transform Infrared (FTIR) spectroscopy as a rapid method for characterizing lignocellulosic biomass. Traditional wet-chemical techniques such as Soxhlet extraction and Klason lignin assays, while accurate, are often slow and labor-intensive. FTIR offers a faster, non-destructive alternative by detecting absorbance peaks associated with specific functional groups like O–H, C=O, and aromatic rings. These spectral features serve as a molecular "fingerprint" that reveals the composition of biomass components, including cellulose, hemicellulose, lignin, and extractives. The research focuses on developing ML models capable of translating FTIR spectra into meaningful compositional and structural information.

The investigation is structured in three phases, each targeting a progressively more focused prediction goal. In the first phase, a full-spectrum multi-output regression model is developed to predict the intensity at every wavenumber (totaling 3551 spectral points) based on nine input features such as biomass category, moisture content, ash, volatile matter, holocellulose, and lignin. Various algorithms—including Partial Least Squares (PLS), Ridge Regression, Random Forest, and Multi-Layer Perceptron (MLP)—are compared for this high-dimensional task.

The second phase shifts focus to broad-range classification. Instead of predicting the exact spectral intensity values, this phase involves identifying whether a significant absorbance peak occurs within predefined spectral intervals (e.g., 3700–3000 cm^{-1} or 1800–1500 cm^{-1}). Here, multi-label classification techniques such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines (SVM) are used to determine the presence or absence of peaks in these regions.

In the third and most targeted phase, the analysis zooms in on narrow spectral intervals such as 3000–2800 cm^{-1} , 1800–1500 cm^{-1} , and 1150–900 cm^{-1} . These ranges are chemically significant, as they correspond to features like aromatic rings in lignin and carbohydrate-related vibrations. Classification models are trained to detect specific absorbance dips within these intervals, directly linking spectral features to key chemical traits.

The study reveals several key findings. In Phase 1, full-spectrum regression proves challenging, with relatively low R^2 values ranging from approximately 0.04 to 0.21. Despite this, the MLP model performs best overall among the algorithms tested. In Phase 2, the task of broad-range peak classification yields better results, achieving Hamming accuracies of up to around 0.75. This improved performance is attributed to the simpler nature of peak detection compared to full spectral prediction. Phase 3 offers the most robust classification results, with Hamming accuracies reaching up to 0.81. Moreover, this approach enhances interpretability, as each narrow spectral band is strongly associated with known chemical features.

Overall, the thesis demonstrates that ML models tailored to different levels of spectral detail—ranging from comprehensive regression to coarse or fine-grained classification—can significantly enhance the utility of FTIR spectroscopy in biomass analysis. The findings support the conclusion that simplified or chemically focused outputs, as developed in Phases 2 and 3, can outperform the more complex full-spectrum predictions of Phase 1. Ultimately, integrating ML with FTIR provides a promising pathway toward rapid, cost-effective, and scalable biomass characterization, with important implications for bioenergy and bio product applications.



BIYOKÜTLE KARAKTERİZASYONU İÇİN FTIR SPEKTRAL PİK NOKTALARININ MAKİNE ÖĞRENMESİ TABANLI TAHMİNİ

ÖZET

Lignoselülozik biyokütlenin yenilenebilir enerji ve biyo-bazlı kimyasallar üretiminde artan önemi, hammadde bileşiminin hızlı, güvenilir ve düşük maliyetli biçimde belirlenmesini zorunlu kılmaktadır. Klasik yaş-kimyasal prosedürler şüphesiz yüksek doğruluk sunar; ancak numune hazırlama, reaktif tüketimi ve saatler ile günler sürebilen deney süresi, endüstriyel ölçekte taramayı yavaşlatır. Fourier Dönüşümlü Kızılötesi (FTIR) spektroskopisi bu darboğazı aşmak için güçlü bir adaydır, çünkü tek bir spektrum ölçümüyle selülozun β -glikozid bağına ait 896 cm^{-1} pikinden ligninin aromatik halkalarının 1510 cm^{-1} titreşimlerine kadar çok sayıda fonksiyonel grup sinyalinin saniyeler içinde yakalar. Bununla birlikte, üst üste binen bantlar, baz çizgisi sapmaları ve partikül boyutu etkileri nedeniyle spektrumun sayısal yorumu hâlen uzman yorumu gerektiren zahmetli bir iştir. Tez çalışması, FTIR verisini makine öğrenmesi (ML) algoritmalarıyla birleştirerek bu yorumu otomatikleştirmeyi ve dolayısıyla biyokütle karakterizasyonunu hem hızlandırmayı hem de nesnel hâle getirmeyi amaçlamıştır.

Araştırmada, 56 farklı biyokütle örneğinin (odunsu kalıntılar, tarımsal atıklar ve enerji bitkileri) kül, nem, uçucu madde, sabit karbon ve holoselüloz gibi dokuz temel özellik üzerinden kategorize edildiği ve her biri için $4000\text{--}600\text{ cm}^{-1}$ aralığını kapsayan 3 551 noktada FTIR transmittans değerinin kaydedildiği deneysel bir veri seti oluşturulmuştur. Bu yüksek boyutlu matris, üç fazlı bir modelleme stratejisine tabi tutulmuştur. İlk fazda “tam spektrum regresyonu” yaklaşımı benimsenmiş, giriş olarak yalnızca dokuz kimyasal/fiziksel parametre kullanılarak spektrumun tüm noktaları aynı anda tahmin edilmeye çalışılmıştır. Çoklu çıkışlı regresyon şeklinde tanımlanan bu sorunda PLS, Ridge, Random Forest ve Çok Katmanlı Algılayıcı (MLP) sinir ağı modelleri test edilmiş; en iyi sonuç MLP ile elde edilmesine karşın R^2 değerlerinin $0,04\text{--}0,21$ aralığında kalması, kısıtlı örnek sayısı ile binlerce çıktı noktasını öngörmenin öngörüldüğü kadar zor olduğunu göstermiştir.

İkinci faz, model hedefini tüm intensiteleri ayrı ayrı üretmekten çıkarıp, sekiz geniş dalgaboyu penceresinde belirgin pik varlığına karar verme problemlerine indirgemıştır. $3700\text{--}3000$, $3000\text{--}2800$, $2800\text{--}1800$, $1800\text{--}1500$, $1500\text{--}1150$, $1150\text{--}900$ ve $900\text{--}450\text{ cm}^{-1}$ gibi aralıklarda “pik mevcut (1) / pik yok (0)” etiketleriyle inşa edilen çok-etiketli sınıflandırma çerçevesinde Lojistik Regresyon, Random Forest, Gradyan Artırma ve RBF Çekirdekli Destek Vektör Makineleri denenmiştir. Kerelemeli çapraz doğrulama sonuçlarına göre Lojistik Regresyon $0,75$ Hamming doğruluğu ve $0,79$ Micro-F1 ile öne çıkarken, Random Forest ve SVM modelleri de $0,68\text{--}0,71$ bandında rekabetçi performans sergilemiştir. Bu bulgu, karmaşık dalga şekillerinin tamamını öngörmek yerine kimyasal olarak anlamlı aralıklarda “pik var/yok” yaklaşımının hem istatistiksel hem de kimyasal açıdan daha tutarlı çıktılar verdiğini göstermiştir.

Üçüncü fazda, biyokütle kompozisyonunu doğrudan yansıtan üç dar pencereye odaklanılmıştır: 3000–2800 cm^{-1} arası alifatik C–H gerilmeleri, 1800–1500 cm^{-1} arası karbonil ve aromatik lignin titreşimleri ve 1150–900 cm^{-1} arası polisakkarit “parmak izi” bölgesi. Bu bölgeler, selüloz/hemiselüloz-lignin dengesini veya pretretman sonrası yapısal değişimleri izlemek için kritik kabul edilir. Dar bantlarda “pik var/yok” sınıflandırması, veri boyutunu ve model karmaşıklığını iyice azaltarak doğruluğu yükseltmiştir. Random Forest modeli bu senaryoda 0,81 Hamming doğruluğu ve 0,89 Micro-F1 ile birinci sıraya yerleşmiş; özellikle lignine özgü 1510 cm^{-1} piki ve hücre duvarı karbonhidratlarını işaret eden 896 cm^{-1} piki neredeyse hatasız tanımlamıştır.

Veri önışlemesi aşamasında baz çizgisi düzeltme, Savitzky-Golay yumuşatma, vektör normalizasyonu ve gerektiğinde ilk türev spektrumlarının hesaplanması gibi adımlar izlenmiş; bunların özellikle sınıf dengesizliği bulunan dar bant modellerinde gürültüyü baskılayarak doğruluğu artırdığı görülmüştür. Ayrıca bazı modellerde polinom özellik genişletme veya değişken önemi temelli dalgaboyu seçimi kullanılmış, böylece giriş boyutunun azaltılmasıyla hem hesaplama süresi kısalmış hem de model genellenebilirliği yükselmiştir. Bu süreç, FTIR spektrumlarını doğrudan ham vektörler olarak değil, kimyasal bilgiyi yoğunlaştıran öznitelik kümeleri olarak ele almanın değerini ortaya koymuştur.

Elde edilen bulgular, FTIR-ML entegrasyonunun üç düzeyde fayda sağladığını ortaya koymaktadır. Birincisi, model eğitildikten sonra yeni bir numunenin spektral paternini saniyeler içinde tahmin edebilmek, laboratuvar işlem süresini katlanarak kısaltır ve yüksek örnek kapasiteli taramalara imkân tanır. İkincisi, makine öğrenmesi insan gözünün kaçırabileceği doğrusal olmayan korelasyonları yakalayarak analitik özneliliği azaltır; örneğin selülozun 1430 cm^{-1} bandındaki küçük bir kaymanın ligninin 1510 cm^{-1} bandındaki zayıf bir artışla birlikte spesifik bir ısı değeri değişimine işaret etmesi gibi karmaşık desenleri keşfedebilir. Üçüncüsü, tam spektrum yerine kimyasal olarak seçilmiş dar bantlarda çalışmak, çıktıların doğrudan yorumlanabilir olmasını sağlayarak proses mühendisleri için hızlı karar desteği sunar; örneğin, 3000–2800 cm^{-1} ’teki metil/asetil sinyallerinin kaybolması buhar patlaması pretretmanının başarıyla lignifikasyonu kıldığını gösterebilir.

Çalışmanın sınırlılıkları da dikkat çekicidir. En önemli kısıt, 56 örnekten oluşan veri setinin hem model karmaşıklığını sınırlaması hem de bazı spectral aralıklarda “pik var” etiketinin çok az gözlemi nedeniyle sınıf dengesizliği yaratmasıdır. Bu durum, belirli aralıklarda yanlış negatiflere yol açabileceğinden, gelecekte sentetik azınlık örnek üretimi (SMOTE) veya sınıf ağırlıklı kayıp fonksiyonlarıyla dengelenebilir. Ayrıca, derin sinir ağları veya dönüştürücü tabanlı modeller ham spektrumu girdi olarak alıp otomatik özellik çıkarımı yaparak özellikle dar bant sınıflandırmalarında daha yüksek doğruluk sağlayabilir; fakat bu modellerin başarılı olabilmesi için daha geniş ve çeşitli bir örnek havuzuna ihtiyaç vardır.

Model şeffaflığı konusu, endüstriyel uygulamalar için güvenilirliğin artırılması açısından öne çıkmaktadır. Bu tezde Random Forest ve Gradyan Artırma modellerinde dalgaboyu bazlı değişken önemi istatistikleri incelenerek 1510, 1240 ve 896 cm^{-1} bölgelerinin sınıflandırma kararlarında kilit rol oynadığı doğrulanmıştır. Gelecekte SHAP (SHapley Additive exPlanations) gibi yorumlanabilir yapay zekâ araçları devreye sokularak modellerin hangi kimyasal sinyalleri ne ölçüde kullandığı netleştirilebilir; böylece “kara kutu” algısı azaltılabilir ve kimyager-mühendis işbirliği teşvik edilebilir.

Sonuçlar literatürle kıyaslandığında, tam spektrum regresyonunda düşük R^2 değerlerinin yaygın olduğu, ancak pik tabanlı sınıflandırmalarda %80 üzeri doğruluklara kolay erişildiği doğrulanmıştır. Örneğin kartal ve Özveren'in benzer çalışmasında da PLS ve MLP modellerinin $R^2 \approx 0,21$ civarında kaldığı, buna karşın dar bant stratejilerinin yüksek isabet sağladığı rapor edilmiştir. Dolayısıyla tezdeki faz geçişi stratejisi—tam spektrumdan geniş aralığa, oradan dar kimyasal bölgeye—yalnızca istatistiksel performansı değil, kimyasal yorumlanabilirliği de sistematik olarak yükseltmiştir.

Pratik açıdan bakıldığında FTIR destekli ML yaklaşımı, biyoyakıt tesislerinde hammadde kabul kontrolünden piroliz, hidrotermal likifaksiyon veya biyokimyasal dönüşüm hatlarında çevrim içi kalifikasyona kadar pek çok noktada gerçek-zamanlı izleme aracı olarak uygulanabilir. Model çıktıları sayesinde selüloz/lignin oranı, uçucu madde miktarı veya olası kül kaynaklı mineralik engeller hızlıca öngörülebilir; proses koşulları bu geribildirimle optimize edilebilir. Ayrıca saha-tabanlı portatif FTIR cihazlarının yaygınlaşması, eğitilmiş ML modellerinin bulut tabanlı sunuculara entegre edilerek tarla veya depo gibi noktalarda anında kompozisyon analizi yapılmasına kapı açacaktır.

Özetle, bu tez, FTIR spektrumlarının makine öğrenmesiyle sentezlenmesinin biyokütle karakterizasyonunda hem hız hem maliyet hem de yorumlanabilirlik boyutlarında çarpıcı avantajlar sunduğunu kanıtlamıştır. Tam spektrum regresyonu, kimyasal ayrıntının tamamını geri kazanmaya imkân tanısa da mevcut veri seti ölçeğinde sınırlı başarı sağlamış; buna karşın geniş ve dar bant sınıflandırmaları, hedeflenmiş bilgi üretimi sayesinde doğruluğu yükseltmiş ve proses mühendislerinin ihtiyaç duyacağı anahtar parametreleri doğrudan sunmuştur. Bu bulgulardan hareketle, model tabanlı FTIR analitiğinin gelecekte yüksek örnek hacimli biyorefinery uygulamalarının standart tanı aracı hâline gelmesi beklenmektedir. Böylelikle biyokütle kaynaklı enerji ve ürün değer zincirlerinde hammadde belirsizliği azalacak, sürdürülebilirlik metriği güçlenecek ve yenilikçi proses tasarımları için sağlam veri temeli sağlanacaktır.



1. INTRODUCTION

1.1 Background of Biomass Characterization

Biomass refers to plant-based organic materials composed largely of lignocellulosic components and minor constituents. The major structural polymers in lignocellulosic biomass are cellulose, hemicellulose, and lignin, along with a variety of extractives (non-structural small molecules) and inorganic ash. In biomass chemistry, holocellulose denotes the total polysaccharide content of the biomass – essentially the sum of cellulose and hemicellulose after removing lignin and extractives (Apaydın Varol & Mutlu, 2023; Segato et al., 2014). Holocellulose typically constitutes over half of woody biomass by weight, indicating a substantial fraction of volatile, carbohydrate-rich material (Apaydın Varol & Mutlu, 2023). Cellulose is a crystalline linear polymer of glucose providing strength to cell walls, whereas hemicellulose is a shorter, branched heteropolymer (e.g. xylans, mannans) that is amorphous and more easily degradable (Esteves et al., 2023). Together, these polysaccharides form the holocellulose fraction which can be enzymatically broken down into sugars or thermochemically converted into biofuels. In contrast, lignin is a complex aromatic polymer comprising phenylpropanoid units (coumaryl, coniferyl, and sinapyl alcohols) linked via ether and carbon–carbon bonds (Segato et al., 2014). Lignin encases the cellulose/hemicellulose matrix in plant cell walls, providing rigidity and resistance to degradation (Segato et al., 2014). Because lignin contains a high proportion of carbon-carbon and aromatic bonds, it tends to have a higher energy content and yields more char upon heating compared to holocellulose. Finally, biomass contains extractives, which are non-structural constituents (often a few percent of dry weight) including fats, resins, sugars, terpenes, and phenolics that can be extracted with solvents (Esteves et al., 2023). These extractives are typically metabolic products not bound in the cell wall and can vary widely between species and seasons (Esteves et al., 2023). Although minor in quantity, extractives can significantly influence biomass properties such as fuel quality and reactivity.

In the context of bioenergy applications, certain bulk properties of biomass are routinely measured to assess its suitability as a fuel or feedstock. These include moisture content, ash content, volatile matter, and fixed carbon, commonly determined via proximate analysis. Moisture is crucial because water does not contribute calorific value; in fact, high moisture significantly reduces the effective heating value of biomass (Demirbas, 2002). Thus, biomass is often dried or pretreated to lower moisture before combustion or pyrolysis. Ash content represents the residual inorganic material after complete combustion. A high ash content is undesirable – it not only lowers the fuel’s heating value but can cause slagging, fouling, and other operational issues in boilers (Demirbas, 2002). Ash in wood is usually low (<1%) comprised mainly of minerals like Ca, K, and Mg, whereas herbaceous or agricultural residues can have higher ash percentages (Esteves et al., 2023). Volatile matter denotes the fraction of biomass that vaporizes and combusts when heated (excluding moisture and carbondioxide). Lignocellulosic biomass generally has a very high volatile matter content (often 70–80% on dry basis) due to the dominance of holocellulose, which decomposes into gases and tars at relatively low temperatures. Indeed, a biomass sample with >55% holocellulose was noted to have a “remarkable volatile fraction” (Apaydın Varol & Mutlu, 2023). In contrast, fixed carbon is the solid carbon left after volatiles are released, largely corresponding to char derived from lignin and other carbon-rich components. For example, biomass with higher lignin content tends to yield more char (higher fixed carbon) during pyrolysis, whereas a holocellulose-rich biomass yields more volatiles (Apaydın Varol & Mutlu, 2023). These relationships mean that the chemical composition (holocellulose vs. lignin ratio, number of extractives, etc.) strongly influences the proximate analysis results and the energy content. Notably, lignin has an inherently higher heating value (~23–26 MJ/kg) than polysaccharides (~18 MJ/kg), and certain extractives can exceed 30 MJ/kg (Esteves et al., 2023). Thus, biomass with high lignin or extractives can have higher calorific value, while biomass with high ash or moisture is energetically less desirable (Demirbas, 2002).

Beyond bulk properties, understanding the chemical composition of biomass (cellulose, hemicellulose, lignin, extractives content) is essential for optimizing biofuel production processes like combustion, pyrolysis, and biochemical conversion. Traditional wet-chemical analysis methods (e.g. Soxhlet extraction for extractives,

Klason method for lignin, etc.) are time-consuming and require significant sample preparation. In this regard, rapid analytical techniques like infrared spectroscopy have become invaluable for biomass characterization. Each type of biomass component contains distinct functional groups that leave “fingerprints” in an infrared spectrum. For instance, holocellulose (cellulose and hemicellulose) is rich in O–H and C–O bonds, whereas lignin contains aromatic rings and various oxygenated functional groups (ethers, carbonyls, etc.). Fourier Transform Infrared (FTIR) spectroscopy captures these signatures as absorption bands at specific wavenumbers, linking chemical structure to measurable spectra. Thus, the complex mixture of holocellulose, lignin, and extractives in biomass can be probed by IR spectroscopy to infer composition and structural features (Apaydin Varol & Mutlu, 2023). For example, O–H stretching vibrations (common to cellulose, hemicellulose, and lignin) give a broad absorption band around 3200–3400 cm^{-1} , and C–H stretching appears near 2900 cm^{-1} in all biomass samples (Apaydin Varol & Mutlu, 2023). More distinctively, the carbonyl (C=O) groups in hemicellulose (e.g. acetyl or uronic ester groups) and in lignin conjugated aldehydes absorb around 1740–1720 cm^{-1} (Zhuang et al., 2020). Lignin, being the only aromatic polymer in biomass, shows prominent aromatic ring vibration bands in the region ~ 1600 –1500 cm^{-1} (Apaydin Varol & Mutlu, 2023; Zhuang et al., 2020). Specifically, aromatic C=C stretching in lignin yields peaks at approximately 1580 cm^{-1} and 1510 cm^{-1} , which are often used as diagnostic markers for lignin content (Zhuang et al., 2020). Cellulose and hemicellulose can be recognized by strong C–O–C and C–O stretching bands in the 1200–1000 cm^{-1} region, associated with the glycosidic bonds and alcohol groups of the polysaccharides (Apaydin Varol & Mutlu, 2023; Zhuang et al., 2020). For instance, the β -(1→4) glycosidic linkages of cellulose give rise to an absorption near 896 cm^{-1} (Zhuang et al., 2020). Many of these bands overlap, but overall the FTIR spectrum of a biomass sample encapsulates its chemical fingerprint: the relative intensities of characteristic peaks reflect the proportions of cellulose/hemicellulose (polysaccharide-associated bands) versus lignin (aromatic bands), plus any unique signals from extractives. In summary, the chemical composition of biomass strongly influences its infrared spectral features, providing a basis for analytical models to correlate spectra with composition and fuel properties.

1.2 Importance of FTIR Analysis in Biomass Research

FTIR has become a cornerstone technique for biomass characterization due to its speed, sensitivity, and minimal sample preparation requirements. FTIR measures the absorbance of infrared light by a sample as a function of wavenumber, producing a spectrum that reflects the sample's molecular bond vibrations. It is especially suitable for lignocellulosic biomass because the major functional groups (O–H, C–H, C–O, C=C, etc.) each absorb at characteristic frequencies, allowing identification of chemical bonds present in cellulose, hemicellulose, lignin, and extractives. Compared to wet-chemical assays, FTIR offers rapid and non-destructive analysis – a spectrum can be obtained in minutes, and modern FTIR instruments using an Attenuated Total Reflectance (ATR) accessory require no complex sample preparation (no pellets or dilutions) (Szymanska-Chargot & Zdunek, 2013). As a result, FTIR is widely used as a screening tool in biomass and biofuel research (Szymanska-Chargot & Zdunek, 2013). For example, researchers routinely employ FTIR-ATR to quickly assess biomass feedstocks for key functional groups or to monitor chemical changes after pretreatment processes. The mid-infrared region ($4000\text{--}400\text{ cm}^{-1}$) is particularly informative: the $1800\text{--}800\text{ cm}^{-1}$ range contains many fingerprint bands unique to biomass components, while $3700\text{--}2800\text{ cm}^{-1}$ covers broad O–H and C–H stretches (Apaydin Varol & Mutlu, 2023; Szymanska-Chargot & Zdunek, 2013). Through reference to established band assignments, one can interpret a biomass FTIR spectrum to qualitatively identify components; for instance, observing a strong peak around 1510 cm^{-1} would indicate aromatic lignin presence, whereas a peak near 1730 cm^{-1} suggests unconjugated carbonyls from hemicellulose or certain extractives (Zhuang et al., 2020). Thus, FTIR provides a molecular fingerprint of the biomass.

However, extracting quantitative or detailed information from FTIR spectra can be challenging due to overlapping peaks, baseline drift, and variations in sample physical state. Data preprocessing is therefore a critical step in FTIR spectral analysis before any interpretation or modeling. Common pre-processing techniques include baseline correction (to remove sloping backgrounds), normalization (to account for concentration or pathlength differences by scaling spectra), smoothing or derivative filtering (to reduce noise and resolve overlapping peaks), and spectral region selection (Mokari et al., 2023). Baseline correction is particularly important for solid biomass ATR spectra, which often show baseline offsets due to scattering or ATR crystal

contact variations (Mokari et al., 2023). Normalizing spectra (e.g., to constant area or to a particular peak) ensures that differences in absorbance are due to compositional changes rather than sample quantity. In addition, transformation methods like taking the first or second derivative of the spectrum can help sharpen peaks and separate broad overlapping bands. Once such preprocessing is applied, multivariate analysis techniques are often used to interpret the complex spectral data. Principal Component Analysis (PCA) is one widely used unsupervised method that reduces the high-dimensional spectral data to a few principal components capturing the majority of variance. Applying PCA to a set of biomasses FTIR spectra can reveal clustering of samples by composition or treatment, and can highlight which wavenumbers contribute most to differences (Szymanska-Chargot & Zdunek, 2013). For example, PCA applied on specific IR regions was able to distinguish different polysaccharide components in plant cell walls, with certain principal component loadings highlighting key bands (e.g. $\sim 1740\text{ cm}^{-1}$ for pectins, $\sim 1370\text{ cm}^{-1}$ for cellulose) (Szymanska-Chargot & Zdunek, 2013). Such analysis demonstrates how specific spectral features correlate with particular structural components. In biomass research, PCA and similar techniques have been used to differentiate wood species, to monitor compositional changes during biomass pretreatment, and to detect contaminants, purely based on spectral patterns.

Overall, FTIR analysis is highly valuable in biomass research because it provides a fast, inexpensive means to characterize chemical composition and structure. The ability to link specific wavenumber bands to functional groups of holocellulose or lignin (as noted earlier) allows researchers to infer properties like cellulose crystallinity or lignin content qualitatively from spectra (Apaydın Varol & Mutlu, 2023; Zhuang et al., 2020). Nevertheless, manual interpretation of spectra has limitations in accuracy and objectivity. This is where computational approaches, including chemometrics and machine learning, become indispensable. By leveraging robust data preprocessing and multivariate analysis, one can develop calibration models – for instance, correlating FTIR spectra to measured biomass composition (from reference chemical analysis) – enabling quantitative predictions. In recent years, the integration of FTIR with advanced data-driven modeling has greatly enhanced its utility, turning FTIR from a primarily qualitative tool into a quantitative predictor for

biomass properties. The next section discusses the motivation for employing machine learning algorithms to interpret FTIR spectral data for biomass characterization.

1.3 Motivation for Machine Learning Applications

While FTIR provides rich spectral information about biomass, interpreting these spectra to obtain quantitative insights (such as exact composition or quality parameters) is non-trivial. Traditional approaches like linear regression or peak-ratio methods often falter given the high dimensionality and collinearity in spectral data. Each FTIR spectrum may consist of thousands of wavenumber intensity values (e.g. 3551 data points for spectra from 4000 to 600 cm^{-1} at 2 cm^{-1} resolution), many of which are correlated. Machine learning (ML) offers a powerful toolkit to handle such complex data, uncover hidden patterns, and build predictive models. The combination of ML algorithms with infrared spectroscopy has been reported as an effective strategy for rapid characterization of biomass and waste materials (Liang et al., 2023). By “learning” from examples – spectra of samples with known properties – supervised ML models can calibrate the relationship between spectral features and those target properties. This enables one to predict unknown sample properties from only the FTIR spectrum, eliminating the need for lengthy laboratory analyses. In the context of biomass, ML models have been used, for example, to predict lignin, cellulose, and holocellulose content from FTIR or NIR (Near-IR) spectra with good accuracy, providing a high-throughput alternative to wet chemistry (Liang et al., 2023). Feature extraction from FTIR data is a key step in this process. Rather than using all spectral points as direct inputs, which can lead to overfitting and obscure the chemistry, one typically distills the data into informative features. This can be done by statistical means (e.g. PCA scores as features) or by selecting specific wavenumbers/bands known to correlate with the property of interest. Recent studies emphasize interpretable feature selection – for instance, choosing a subset of “high-loading” spectral peaks that have known physicochemical relevance (like the 1508 cm^{-1} lignin band or 896 cm^{-1} cellulose band) – and using those as inputs to ML models (Liang et al., 2023). Such approaches marry domain knowledge with data-driven modeling, yielding models that not only perform well but are easier to interpret chemically (Liang et al., 2023). In our work, we similarly extract meaningful features from the raw FTIR spectra (through careful preprocessing and selection of significant wavenumber

regions) to feed into ML algorithms, thereby focusing the models on the most relevant spectral variations.

It is important to distinguish between supervised and unsupervised machine learning approaches in this context. Unsupervised techniques like PCA or cluster analysis do not use external labels; they aim to find natural groupings or patterns in the spectral data. These can be useful for exploring dataset structure (for example, clustering spectra by biomass type or treatment without prior knowledge) and for dimensionality reduction. In contrast, supervised learning involves training models on spectra that are labeled with known outputs (responses). In biomass FTIR analysis, common supervised tasks include regression (predicting a continuous variable such as % lignin or heating value from the spectrum) and classification (identifying the class of a biomass sample, e.g. softwood vs. hardwood, based on its spectrum). Supervised learning algorithms range from relatively simple multivariate regression methods like PLS – long used in chemometrics – to more complex methods like Support Vector Machines, Random Forests, neural networks, and ensemble methods. These algorithms can capture non-linear relationships and interactions between spectral features and target properties that manual analysis could miss. For example, a machine learning model can learn a combination of absorbance changes across multiple wavenumbers that together predict a property, even if no single peak is uniquely indicative. This ability to handle multivariate correlations is crucial for biomass spectra, where peaks overlap and baseline shifts occur. Model evaluation is then performed to ensure the predictive reliability of these ML models. For regression models, typical performance metrics include the RMSE – which quantifies the average prediction error in the same units as the target – and the coefficient of determination (R^2) – which indicates the proportion of variance in the target explained by the model (Mokari et al., 2023). Lower RMSE and higher R^2 signify better predictive accuracy and fit. For classification models, evaluation uses metrics such as overall accuracy (the fraction of correctly classified instances) and analysis of the confusion matrix which yields metrics like precision, recall (sensitivity), specificity, and the F_1 -score (Mokari et al., 2023). These metrics provide insights into how well the model distinguishes different classes (for instance, how accurately a model can discriminate between different biomass feedstock types or the presence/absence of a certain component). In our research, we employ both regression and classification paradigms; thus, we report

metrics appropriate to each (e.g. RMSE and R^2 for predicting spectral intensities or concentrations, and accuracy and confusion matrix-derived measures for predicting categorical outcomes). By analyzing these metrics – and comparing them across different modeling approaches – we can assess the effectiveness of integrating machine learning with FTIR data.

The motivation for applying machine learning to FTIR-based biomass analysis is multifold. First, ML can dramatically accelerate biomass characterization. Once a model is trained, obtaining a prediction for a new sample's composition or quality takes only seconds after measuring its FTIR spectrum. This is invaluable for high-throughput screening in bioenergy research, where hundreds of biomass samples may need to be evaluated. Second, ML models can potentially improve accuracy and objectivity in analysis by learning from large datasets, whereas human interpretation of spectra might overlook subtle features or suffer from bias. Third, ML enables the detection of complex spectral patterns associated with chemical properties that are not captured by simple peak height comparisons. For example, a random forest or neural network might learn that a slight shift in a cellulose band combined with a change in a lignin band is predictive of a certain pretreatment effect – relationships that would be hard to deduce by eye. Finally, the integration of FTIR with ML aligns with the broader trend of developing rapid, in-line analytical techniques for biofuel production. It provides a strong methodological advantage: non-destructive, real-time analysis with predictive insight, reducing the need for costly and time-consuming wet chemistry for every sample. In summary, machine learning augments FTIR spectroscopy, enabling it to serve not just as a qualitative diagnostic tool but as a quantitative predictive platform for biomass characterization (Liang et al., 2023). This thesis builds upon that motivation, aiming to demonstrate that carefully developed ML models can reliably interpret FTIR spectra of biomass to yield meaningful compositional and structural information.

1.4 Research Objectives and Hypotheses

This research aims to develop and validate a ML framework that leverages FTIR spectroscopy data of lignocellulosic biomass to extract meaningful chemical insights. The approach follows a structured, three-phase plan. In Phase 1, the objective is to predict the complete FTIR spectral profile using ML techniques. Phase 2 moves

toward identifying broader wavenumber intervals, such as 4000–3700, 3700–3000, 3000–2800, 2800–1800, 1800–1500, 1500–1150, 1150–900, and 900–450 cm^{-1} . Finally, Phase 3 focuses on narrow, chemically specific regions, particularly 3000–2800, 1800–1500, and 1150–900 cm^{-1} , where key functional groups are most likely to appear.

An essential aspect of this study involves examining how data preprocessing and feature selection techniques—such as baseline correction, normalization, and waveband selection using principal components or specific intervals—impact the performance and robustness of ML models. The underlying assumption is that by cleaning and compressing the spectral data, and using domain knowledge to bin wavelengths effectively, the resulting models will be more accurate and easier to interpret.

The research also seeks to compare regression-based and classification-based ML strategies. For example, multi-output regression may be best suited for full-spectrum predictions, while binary classification could be more appropriate for identifying the presence or absence of signals within specific spectral intervals. Each method will be evaluated on how effectively it manages the complexity of broad versus narrow IR spectral regions.

Another core objective is to establish the feasibility and advantages of ML-enhanced FTIR analysis as a faster, potentially more cost-effective alternative to traditional chemical assays. By training models to quickly interpret biomass spectra, the workflow could significantly speed up the screening of bioenergy feedstocks, enabling quicker and more informed decisions.

The study is grounded in several hypotheses. First, it is expected that FTIR spectra contain sufficient chemical information to support accurate ML predictions across all three phases, capturing distinctive transmittance dips that reveal biomass composition. Second, preprocessing and feature selection are believed to be crucial; unprocessed spectra may carry excessive noise or redundancy, whereas thoughtful preparation of the data will enhance both accuracy and clarity. Third, a phased modeling strategy is hypothesized to improve both interpretability and focus. While full-spectrum modeling offers a comprehensive view, broad-interval classification helps to localize major functional group regions, and narrow-band targeting allows for the detection of

specific chemical bonds such as aromatic, aliphatic, or carbohydrate signals. Lastly, it is proposed that ML-FTIR models, once trained, will be capable of replacing some wet-lab procedures, offering practical utility in real-world biomass analysis through rapid and accurate predictions.

Altogether, this research undertakes a detailed exploration of FTIR-ML integration through a three-phase methodology, with the expectation that model specificity and chemical interpretability will improve at each step. The final goal is to establish ML-enhanced FTIR as a valuable, efficient tool for characterizing lignocellulosic biomass in the context of bioenergy and beyond.

1.5 Scope of the Study

This thesis unfolds through three progressive phases of investigation, each deepening the integration of machine learning (ML) with FTIR spectroscopy, and refining both the prediction target and the focus of the modeling effort. The scope begins broadly, aiming for full-spectrum reconstruction, then shifts toward interval-based classification, and finally concentrates on narrow, chemically significant regions.

In the first phase, the task is to predict all 3551 wavenumber intensities within the mid-infrared (mid-IR) range of FTIR spectra. This phase is framed as a multi-output regression problem where the ML model attempts to reconstruct the entire spectral profile of a biomass sample, potentially using simpler analytical measurements or partial spectral data as input. Serving as a high-complexity benchmark, this phase challenges the model to learn the intricate spectral patterns characteristic of various biomass types. The high dimensionality and inherent noise in spectral data make this task particularly demanding, laying the groundwork for the more focused strategies pursued in later phases.

The second phase transitions from exhaustive spectral reproduction to a more generalized classification of broader spectral regions. Instead of predicting every single wavenumber intensity, the model assesses whether prominent transmittance dips (or absorbance peaks) are present within predefined spectral bins. These bins include intervals such as 4000–3700 cm^{-1} , 3700–3000 cm^{-1} , 3000–2800 cm^{-1} , 2800–1800 cm^{-1} , 1800–1500 cm^{-1} , 1500–1150 cm^{-1} , 1150–900 cm^{-1} , and 900–450 cm^{-1} . Each region corresponds to key functional group signatures, such as O–H stretching

above 3000 cm^{-1} , carbonyl and aromatic absorptions near $1800\text{--}1500\text{ cm}^{-1}$, and carbohydrate-related features in the $1150\text{--}900\text{ cm}^{-1}$ range. The model classifies each bin as either "peak present" or "absent," simplifying the output into aggregated spectral descriptors. This coarse classification enables quicker and often more reliable identification of functional group signals, which is advantageous for applications like lignin detection or rapid biomass screening. Supervised classification algorithms—such as logistic regression, support vector machines, random forests, and gradient boosting—are employed to detect these broad spectral features, and the reduced granularity is expected to yield higher model accuracy than the more complex task of Phase 1.

In the third phase, attention narrows to a few critical spectral intervals: $3000\text{--}2800\text{ cm}^{-1}$, $1800\text{--}1500\text{ cm}^{-1}$, and $1150\text{--}900\text{ cm}^{-1}$. These regions are selected based on their well-documented associations with specific chemical functionalities, including aliphatic C–H stretching, carbonyl and aromatic ring absorptions, and the complex carbohydrate fingerprint region. The focus here is to build models that offer high chemical interpretability, isolating absorbance peak shapes or presence within these specific ranges. Depending on the objective, either regression (to predict curve shapes) or classification (to detect peak presence) may be applied. This phase represents a hybrid of the previous two: it retains Phase 1's spectral detail but limits the scope to chemically important targets, similar to Phase 2's interpretive clarity. It also enables comparative analysis across the selected intervals, identifying which of them are most predictable based on the rest of the available spectral data.

Across all three phases, the complexity of the ML task is systematically varied. Phase 1 establishes a demanding and comprehensive baseline, Phase 2 simplifies the problem by aggregating the data into functional group-level bins, and Phase 3 concentrates on the most chemically informative regions for fine-tuned analysis. The study is limited to FTIR spectral data from lignocellulosic biomass and does not involve the development of new spectroscopic theories. However, established knowledge of spectral band assignments is heavily used to interpret the chemical relevance of model outputs. This phased approach allows for a strategic exploration of FTIR–ML integration, optimizing both interpretability and predictive performance in the context of biomass characterization.

1.6 Thesis Structure

This thesis follows a structured progression through five main chapters, each designed to reflect a logical development of the research — from establishing foundational context, reviewing existing literature, detailing the methods employed, presenting and discussing results, and finally summarizing key conclusions and proposing avenues for future work.

The first chapter, Introduction, introduces the background and motivation for the study, focusing on biomass characterization and the potential of FTIR spectroscopy as a rapid analytical technique. It defines core concepts such as lignocellulosic composition, holocellulose, and lignin, while emphasizing the need for advanced machine learning approaches to fully leverage spectral data. The research objectives and hypotheses are clearly articulated, followed by a detailed description of the study's scope and structure.

The Chapter 2: Literature review delves into two interconnected fields. The first section explores the chemical makeup of lignocellulosic biomass and the relevance of FTIR spectroscopy in identifying key functional groups like aromatic rings, carbonyls, and polysaccharides. The second section reviews how machine learning and chemometric methods have been applied to infrared spectral analysis. Particular attention is given to multi-output regression, classification strategies for spectral features, and preprocessing techniques such as baseline correction and normalization. This review identifies current gaps and highlights how ML could enhance interpretation and prediction of spectral data, forming the theoretical basis for the modeling strategies employed in this thesis.

The third chapter, Materials and methods, details the experimental and computational workflow of the research. It begins with a description of the biomass samples, including sourcing, preparation, and any accompanying reference analyses. The FTIR data acquisition process is then presented, covering instrument parameters, spectral resolution, and the preprocessing steps applied to the raw spectra. The chapter then outlines the machine learning strategies employed across three phases: Phase 1 involves full-spectrum multi-output regression; Phase 2 focuses on classifying broad wavenumber intervals where significant absorbance peaks may occur; and Phase 3 narrows the analysis to specific chemically relevant regions (such as 3000–2800,

1800–1500, and 1150–900 cm^{-1}). Finally, the machine learning models used—including PLS, logistic regression, support vector machines, random forest, and gradient boosting—are introduced, along with details on hyperparameter tuning, cross-validation, and feature selection methods like principal component analysis.

The Chapter 4: Results and discussion presents and interprets the outcomes of the three-phase modeling approach. For Phase 1, the chapter details the model performance in predicting full-spectrum intensities, including metrics such as RMSE and R^2 , and discusses the regions of the spectrum where the model performs well or struggles. In Phase 2, the results of classifying absorbance peaks in broad spectral intervals are analyzed using classification metrics like accuracy, F1-score, and confusion matrices. This section also compares these results with Phase 1 to evaluate whether reducing output complexity enhances robustness and interpretability. Phase 3 results are then presented, focusing on the narrow spectral regions most associated with lignin and carbohydrate signals. Performance is again evaluated using both regression and classification metrics, and the interpretive value of each targeted spectral window is discussed in terms of chemical specificity. Throughout this chapter, model comparisons and practical implications for biomass screening and analysis are highlighted.

The final chapter, Conclusion and future work, synthesizes the key findings from each modeling phase and reflects on the research objectives. It discusses the effectiveness of ML-FTIR integration in providing rapid, data-driven insights into biomass composition and highlights the trade-offs between model complexity and interpretability. The chapter also considers broader implications, such as potential applications in real-time process monitoring or large-scale biomass screening. Finally, it addresses limitations of the current study and outlines future research directions, including expanding the dataset, refining spectral binning strategies, and exploring other spectroscopic techniques or deep learning models.

Together, these chapters form a coherent narrative that guides the reader through a multi-phase exploration of FTIR-based machine learning, demonstrating how increasingly focused predictive models can enhance chemical interpretation and support practical biomass characterization.



2. LITERATURE REVIEW

2.1 Biomass Composition and Analysis Techniques

Biomass, particularly lignocellulosic biomass, is primarily composed of three structural components: cellulose, hemicellulose, and lignin (Tayyab et al., 2018). Cellulose is a fibrous, crystalline polysaccharide (a polymer of glucose) that provides tensile strength to plant cell walls. Hemicellulose is a shorter, branched heteropolysaccharide that, together with lignin, fills the amorphous space around cellulose fibers. Lignin is a complex aromatic polymer that binds to and stiffens the cell wall, giving compressive strength and resistance to degradation (Jesus et al., 2024). Quantitatively, cellulose typically constitutes about 40–50% of dry lignocellulosic biomass, hemicellulose about 20–30%, and lignin roughly 15–25%, though these ratios vary with biomass type (e.g., wood vs. grass) (Tayyab et al., 2018). In addition to these main fractions, biomass contains smaller amounts of extractives and inorganic minerals. Extractives encompass various organic compounds not bound in the cell wall (such as resins, oils, fats, waxes, and phenolics) and can make up to ~5–15% of biomass; they are responsible for properties like color, odor, or decay resistance in wood (Jesus et al., 2024). The inorganic component, reported as ash content, includes mineral salts (e.g., silica, calcium, potassium); in wood it is often low (around 0.5% or less) (Jesus et al., 2024) but can be higher in agricultural residues. Biomass also inherently contains moisture (water content) whose level depends on storage and ambient conditions; moisture does not contribute to dry-matter composition but is important for handling and is usually measured separately. The sum of cellulose and hemicellulose is sometimes termed holocellulose, representing the total carbohydrate fraction of the biomass (Javier-Astete et al., 2021). In essence, the quality and utility of a biomass feedstock (for biofuel, materials, etc.) are largely determined by these components – for example, high holocellulose is desirable for fermentable sugar production, whereas high lignin can impede enzymatic conversion but is useful for combustion energy.

2.1.1 Conventional analysis methods

Determining the composition of biomass has traditionally involved a series of wet-chemical and gravimetric techniques. Standard protocols exist to quantify each fraction: for instance, cellulose and hemicellulose can be measured by hydrolyzing the biomass in acid and analyzing the released sugars (e.g., via high-performance liquid chromatography), while lignin is often measured as the residue after sulfuric acid digestion (known as Klason lignin) along with UV spectroscopy for acid-soluble lignin. Extractives are determined by solvent extraction (such as Soxhlet extraction with ethanol/benzene or dichloromethane) that removes resins, waxes, and oils prior to structural analysis. Ash content is obtained by igniting the biomass at high temperature (≥ 575 °C) and weighing the inorganic residue, and moisture is measured by oven-drying samples at 105 °C to constant weight. These conventional techniques are time-consuming and labor-intensive, requiring significant sample preparation and specialized chemical reagents. For example, a full compositional analysis might involve successive extraction, acid hydrolysis, filtration, and titration steps which can take several days for a batch of samples. While the results are accurate, the throughput is low and not easily scalable for large sample numbers or real-time analysis. In practice, there is often interest in faster analytical methods to screen biomass, especially in bioenergy research where hundreds of samples may need evaluation. This has led to exploration of spectroscopic techniques as rapid alternatives (Javier-Astete et al., 2021). FTIR, in particular, has gained traction as a complementary method because it can infer chemical composition from a quick, non-destructive measurement of the sample's infrared spectrum (Javier-Astete et al., 2021).

2.1.2 FTIR spectroscopy for biomass characterization

FTIR spectroscopy is an analytical technique that measures how a sample absorbs light across the mid-infrared range of wavelengths. The resulting FTIR spectrum – essentially a plot of absorbance (or transmittance) versus wavelength (typically reported as wavenumbers, cm^{-1}) – provides a fingerprint of the sample's molecular bonds. Biomass contains a variety of functional groups (O–H, C–H, C=O, C–O, aromatic rings, etc.) associated with its components, and each of these groups absorbs IR light at characteristic frequencies. Table 2.1 (in the context of literature) or various studies document the typical band assignments for biomass: for example, broad O–H

stretching around 3300 cm^{-1} (due to hydroxyl groups in cellulose and hemicellulose), C–H stretching of methyl and methylene groups near 2900 cm^{-1} , and a series of bands in the fingerprint region ($1800\text{--}800\text{ cm}^{-1}$) that correspond to the core functional groups of the biomass polymers (Javier-Astete et al., 2021). Notably, the carbonyl (C=O) stretch around $1730\text{--}1740\text{ cm}^{-1}$ is often attributed to acetyl and uronic ester groups in hemicellulose or to certain esterified extractives, the aromatic ring vibrations of lignin appear near 1600 cm^{-1} and 1510 cm^{-1} (skeletal vibrations of the benzene rings in lignin) (Javier-Astete et al., 2021), and C–O stretching coupled with C–H bending in polysaccharides gives strong signals in the $1050\text{--}1150\text{ cm}^{-1}$ range (dominated by cellulose and hemicellulose) (Javier-Astete et al., 2021). For instance, an FTIR spectrum of wood typically shows a lignin-associated peak at $\sim 1515\text{ cm}^{-1}$ (aromatic ring vibration) and carbohydrate-associated peaks around 1375 , 1155 , 1050 , and 898 cm^{-1} (various cellulose and hemicellulose vibrations, including the β -glycosidic linkage vibration near 898 cm^{-1}) (Javier-Astete et al., 2021). These spectral features allow qualitative identification of biomass constituents: one can often tell if a sample has a higher lignin content by the relative intensity of the aromatic bands, or detect the presence of certain extractives by peaks (for example, a sharp peak around 1700 cm^{-1} might indicate carbonyl-containing extractives or hemicellulose acetyl groups).

One clear advantage of FTIR in biomass analysis is speed and minimal sample preparation. Using an ATR accessory (commonly used for solid biomass samples), one can often analyze a ground biomass sample in a matter of minutes or less, with no chemical reagents – far quicker and simpler than traditional wet-chemistry assays (Javier-Astete et al., 2021). FTIR is also non-destructive or only mildly destructive (the sample remains largely intact except for drying and pressing against the ATR crystal), meaning the same sample can be preserved for other analyses if needed. Because of these benefits, FTIR has been widely adopted as a screening tool to estimate biomass composition in research and industry (Javier-Astete et al., 2021; Zhuang et al., 2020). Studies have demonstrated that FTIR spectral data correlates with contents of cellulose, hemicellulose, lignin, and even minor components, enabling it to identify or quantify these constituents indirectly (Javier-Astete et al., 2021). For example, Javier-Astete et al. (2021) note that FTIR-ATR spectroscopy has been successfully used to identify major wood components – “cellulose, hemicellulose, lignin, monosaccharides, extractive compounds and proteins” – in various forest

species when coupled with suitable data analysis (Javier-Astete et al., 2021). In practical terms, an FTIR spectrum encapsulates the composite signal of all components, and by using reference samples with known composition, one can develop calibration models to predict composition from spectra. Thus, FTIR offers a rapid, molecular fingerprinting approach to biomass characterization, providing insight into chemical makeup without the need to perform each chemical test separately.

However, interpreting an FTIR spectrum of a complex mixture like biomass is not straightforward by simple inspection. Many absorption bands overlap or are broadened due to the heterogeneous nature of lignocellulosic materials (Javier-Astete et al., 2021). For instance, both cellulose and lignin contribute in the region around 1100–1000 cm^{-1} , and both hemicellulose and extractives can have C=O bands near 1730 cm^{-1} (Javier-Astete et al., 2021). This overlap means that a single peak usually cannot be unambiguously assigned to one component's quantity. Instead, the overall pattern must be analyzed in a multivariate way. As a result, the role of FTIR in quantitative biomass analysis is typically realized in combination with computational methods: the raw spectral data are subjected to chemometric or machine learning algorithms that can deconvolute the contributions of components. In short, FTIR provides the data-rich spectra, and advanced data analysis links those spectra to biomass composition. This approach – using multivariate calibration models to relate spectral features to laboratory-measured composition – has been a cornerstone of modern biomass analysis and is discussed in the next section (Javier-Astete et al., 2021). In summary, FTIR-based characterization of biomass leverages the fact that spectral signatures correlate with chemical composition; with proper models, one can predict quantities like holocellulose or lignin content from an FTIR scan, offering a powerful alternative or complement to classical wet chemistry methods (Javier-Astete et al., 2021).

2.2 Machine Learning for Spectral Data

The complex nature of FTIR spectra for biomass – containing hundreds of data points (wavenumbers) with overlapping signals – necessitates the use of multivariate data analysis and machine learning (ML) techniques. Machine learning in this context refers to a broad class of algorithms capable of finding patterns or relationships in data,

which includes traditional chemometric methods as well as modern statistical learning approaches. These methods are crucial for extracting quantitative or categorical information from spectra that are impossible to interpret by simple univariate peak analysis. In recent years, ML models have become integral to spectroscopic analysis across disciplines, enabling rapid predictions once calibrated (Fadlelmoula et al., 2023). This section provides an overview of the ML models commonly applied to FTIR (and other spectroscopic) data, their basic principles (in a non-technical way), and examples of their use in biomass characterization.

2.2.1 Multivariate regression models

A primary application of ML in FTIR analysis is to perform regression, i.e. predict a continuous parameter (such as percentage of lignin or cellulose) from the spectrum. The most widely used regression approach in infrared spectroscopy is PLS regression, which originates from chemometrics. PLS is well-suited to spectroscopic data because it can handle highly collinear and high-dimensional predictors. In essence, PLS works by projecting the original spectral variables into a smaller set of latent factors (sometimes called PLS components) that are linear combinations of the original wavelengths, chosen to maximize the covariance with the property of interest (the response variable) (Javier-Astete et al., 2021). By doing so, PLS reduces noise and redundancy, focusing on the spectral variations that most strongly relate to the target (e.g., lignin content). PLS regression has been a cornerstone in building calibration models for instruments like FTIR and NIR spectrometers (Javier-Astete et al., 2021). For example, researchers have used PLS to correlate FTIR spectra with the chemical composition of wood and achieved good predictive performance, effectively substituting for laboratory tests (Javier-Astete et al., 2021). In one study, FTIR-PLS models built on wood samples could predict cellulose, hemicellulose, and lignin concentrations with reasonable accuracy (with prediction errors on the order of 1–2% for major components), demonstrating that PLS could capture the relevant spectral information for each constituent (Javier-Astete et al., 2021). Another study on forest biomass residues used first-derivative FTIR spectra with PLS and found strong correlations ($R^2 > 0.8$) between the spectra and measured properties (Acquah et al., 2016a). In that case, the models predicted lignin and extractives particularly well, with performance metrics indicating that the spectroscopic method could reliably estimate those fractions (Acquah et al., 2016a). These examples underscore PLS's value as a

baseline method for spectral regression – it is often the first tool applied due to its effectiveness and the fact that it provides a sort of built-in feature extraction (via the latent variables).

Beyond PLS, a variety of other regression techniques have been applied to spectral data to improve or complement the results. Traditional multiple linear regression is generally not used directly on full spectra (due to severe multicollinearity and overfitting risk), but modern machine learning regressors can handle complex data patterns. Support Vector Regression (SVR), for instance, is the regression variant of Support Vector Machines; it fits a relationship by finding a function (potentially nonlinear via kernel transformations) that has at most a certain error for all training points and is as flat as possible (maximizing margin). SVR can model nonlinear trends in spectra, such as subtle shifts in peak shapes with composition, which a linear PLS might not capture. Ensemble methods like RF regression have also been explored. Random Forests consist of many decision tree models voting together; each tree partitions the spectral feature space based on thresholds of absorbance at certain wavelengths, and the ensemble average improves generalization. RF is quite robust to overfitting and can naturally model interactions and nonlinear effects in spectral data. For example, in one bioenergy study, a random forest model outperformed a neural network in predicting yields of bio-oil, char, and gas from biomass based on input features (Li et al., 2023), indicating the strong performance of ensemble methods for complex prediction tasks. Although that example pertains to thermochemical conversion outputs, the principle carries to spectral analysis: RF can be effective in cases where the relationship between absorbance and concentration is nonlinear or when there are important interactions between different spectral regions.

Another class of regression models gaining traction is neural networks, particularly artificial neural networks (ANNs) with one or more hidden layers. These models can approximate complicated functional relationships by learning weights in a network of interconnected nodes (neurons). In spectral analysis, ANNs have been used to relate absorbance values (or transformed features like principal components) to chemical properties. They can capture nonlinearities and intricacies that linear methods might miss, given sufficient training data. A recent study provides an illustrative comparison: Pushpa et al. (2024) developed infrared-based ML models for predicting multi-feedstock lignocellulosic composition and found that an optimized ANN model

outperformed PLSR models in prediction accuracy (Pushpa et al., 2024). The ANN was better able to fit the calibration data across a diverse set of biomass types, improving the quantification of cellulose, hemicellulose, and lignin. This result suggests that when the goal is highly accurate prediction (and if a robust dataset is available), more complex ML models like ANNs can add value beyond the classical PLS approach. That said, neural networks require careful tuning (e.g., architecture, regularization) and are often viewed as "data-hungry" – they typically need a larger number of training samples to learn reliably, to avoid overfitting noise in the spectra. In practice, the choice of regression model often involves a trade-off between model complexity and the amount of calibration data available. Simpler models (like PLS or ridge regression) may perform as well as more complex ones when data are limited, whereas complex models can excel with more data and variability.

2.2.2 Classification models

In addition to predicting continuous composition values, machine learning is used with FTIR data for classification tasks in biomass research. Classification involves predicting category labels – for example, identifying the species of an unknown wood sample, determining if biomass has been pretreated or not, or categorizing a sample by quality grade. One common approach for spectral classification is to use discriminant analysis. Linear Discriminant Analysis (LDA) or its close relative Partial Least Squares-Discriminant Analysis (PLS-DA) are techniques that find a linear combination of features (wavenumber intensities or their derivatives) that best separates two or more classes. PLS-DA is essentially the application of PLS to a binary or multi-class label (with the PLS regression trying to predict class membership), resulting in a predictive discriminant model. These methods have been effectively used for wood and biomass identification. For instance, Pasquini et al. demonstrated that three Amazonian wood species could be successfully discriminated by FTIR combined with PLS-DA, achieving classification accuracies above 91% using key spectral peaks associated with cellulose, lignin, and hemicellulose (Jesus et al., 2024). In that study, the distinct IR signatures (such as differences in lignin aromatic band intensity and carbohydrate region shapes) provided enough information for the model to differentiate species with high confidence. Similarly, FTIR-PLSDA models have been used to distinguish compression wood vs. normal wood in pine based on lignin band

differences (Jesus et al., 2024), indicating that even subtle anatomical or growth differences in biomass can be detected via spectral patterns.

More advanced or non-linear classifiers have also been applied. SVM in classification mode are popular due to their ability to handle high-dimensional data and create complex decision boundaries. An SVM classifier finds the optimal hyperplane that separates classes by maximizing the margin between class clusters in a transformed feature space (using kernel functions to allow nonlinear separation). In the context of biomass, Souza et al. (2024) (published in RSC Advances) recently showed that combining PCA with an SVM classifier on FTIR data enabled accurate identification of different Eucalyptus wood species (Jesus et al., 2024). In their approach, PCA was first used to reduce the dimensionality of the FTIR spectra and capture the major variance, then the top principal components were fed into an SVM which classified the species. This yielded a practical method for wood species identification that could aid in quality control and prevent species fraud in the timber industry (Jesus et al., 2024). Other studies have also used hierarchical cluster analysis (HCA) and principal components to group samples, followed by discriminant analysis (like LDA) to formalize classification (Jesus et al., 2024). For example, FTIR spectra from a mix of hardwood and softwood samples were first clustered by HCA into groups, and then PCA-LDA was used to successfully identify the geographic origin of the wood samples (distinguishing woods from different growing locations by their spectral fingerprints) (Jesus et al., 2024). These multistep approaches illustrate how unsupervised methods (clustering, PCA) can be combined with supervised classification to tackle complex categorization problems.

2.2.3 Unsupervised feature extraction

Even when the ultimate goal is regression or classification, unsupervised learning methods like PCA play a key supporting role in spectral data analysis. PCA reduces the dimensionality of spectral data by finding a set of new orthogonal axes (principal components) that capture the greatest variance in the data (Fadlelmoula et al., 2023). By projecting spectra into the space of the first few principal components, one can both simplify the dataset and often visualize natural groupings or trends. In biomass FTIR analysis, PCA is commonly used to observe how samples cluster according to composition or type. For instance, PCA might reveal that samples separate along one

principal component according to lignin content (since that component's loading is heavily weighted on lignin-associated peaks), while another component might separate samples by a different attribute (like one species vs another, if spectral differences exist) (Jesus et al., 2024). PCA by itself can thus provide an initial check on whether spectral differences correspond to meaningful chemical or class differences. Moreover, the principal components (or other features derived from them) are often used as inputs to subsequent ML models to avoid overfitting and to improve robustness. This approach, known as feature extraction, was highlighted in the examples above (e.g., PCA+LDA or PCA+SVM), and is widely recommended when the number of spectral variables is large compared to the number of samples. It effectively distills the data, often reducing noise by ignoring minor variance that could be due to measurement artifacts.

2.2.4 Data preprocessing and other techniques

Before applying ML models, FTIR spectral data typically undergo preprocessing steps which can be considered part of the analytical pipeline. Common preprocessing includes baseline correction (to remove any sloping background in the spectrum), normalization (such as unit vector normalization or standard normal variate correction (Pushpa et al., 2024) to account for path length or concentration differences), and derivative spectroscopy (calculating first or second derivatives of the spectral curve to sharpen peaks and resolve overlapping signals). For example, using a first-derivative of the FTIR spectrum can enhance subtle features and was done by Acquah et al. to improve PLS model performance for forest residues (Acquah et al., 2016a). Smoothing filters like the Savitzky–Golay filter are also applied to reduce high-frequency noise while preserving peak shape. The choice of preprocessing can significantly impact the subsequent ML model – a well-chosen preprocessing can make the difference between a successful calibration and a failed one. Indeed, researchers often test multiple preprocessing schemes (e.g., combinations of derivative + normalization) and select the one yielding the best predictive model in cross-validation (Javier-Astete et al., 2021). In one study, an automated tool was used to evaluate various spectral pretreatments on FTIR data to optimize the prediction of each component (Javier-Astete et al., 2021), underscoring that this is an important empirical step.

After preprocessing, feature selection may be employed to reduce the spectral variables to those most informative for the task. Instead of using all wavelengths from, say, 4000 to 600 cm^{-1} , one might select specific regions known to contain relevant signals (like the fingerprint region 1800–800 cm^{-1}). There is evidence that focusing on such informative regions can improve model performance. For example, Zhang et al. (2020) reported that PLS regression models restricted to key sub-intervals of the spectrum slightly outperformed those built on full-range spectra for predicting cellulose, hemicellulose, and lignin in biofuel pellets (He et al., 2022a). The 1000–1800 cm^{-1} range, rich in lignocellulosic signatures, provided better signal-to-noise by excluding areas like 2000–2700 cm^{-1} which contained little useful information (He et al., 2022a). This kind of interval selection or variable selection can be done via algorithms as well (e.g., Genetic Algorithms, interval PLS, or based on variable importance metrics from an initial model). Machine learning models like Random Forest inherently give variable importance scores by measuring how much each wavelength contributes to reducing prediction error in the trees. Such information can be used to trim down the input features to the most predictive wavelengths, simplifying the model and sometimes improving generalization. The overall goal of these steps is to ensure that the model builds its predictions on real chemical signal rather than noise or artifacts.

2.2.5 Illustrative applications in literature

Numerous studies have integrated ML with FTIR for biomass characterization, validating the approach. To highlight a few: Acquah et al. (2016) developed PLS regression models using FTIR reflectance spectra to predict the composition and fuel properties of forest logging residues. They reported that the FTIR-PLS models could predict lignin and extractives with higher accuracy ($R^2 > 0.80$) than cellulose, and also accurately estimated thermochemical parameters like volatile matter and fixed carbon content (with RPD > 2 , indicating useful predictive ability) (Acquah et al., 2016a). This demonstrated that FTIR coupled with ML could rapidly provide information relevant to bioenergy conversion, information that would traditionally require separate chemical and thermal analyses. Another study on wood pellets by Liu et al. (2019) combined FTIR with PLS and achieved calibration R^2 values around 0.95 for cellulose, hemicellulose, and lignin prediction, with excellent stability in cross-validation (He et al., 2022a). The strong correlation between spectral data and composition in that work

confirms that mid-IR spectra contain sufficient quantitative information when processed with robust multivariate models. On the classification side, de Oliveira et al. (2024) used FTIR plus multivariate analysis to identify five Brazilian wood species, obtaining a high classification accuracy by selecting appropriate spectral ranges and using relatively simple algorithms for discrimination (Jesus et al., 2024). They noted that even though the FTIR spectra of the species were very similar (due to all being lignocellulosic), subtle but consistent differences could be captured by the model to differentiate each species (Jesus et al., 2024). These prior works collectively demonstrate that ML-driven spectral analysis can address a variety of biomass characterization needs – from determining chemical composition to recognizing material identity – with speed and accuracy. They provide a foundation for further advances, while also indicating certain limitations (in cases where models struggled, such as predicting one component less accurately, or requiring careful selection of spectral features for success).

In summary, machine learning techniques have become indispensable for unlocking the potential of FTIR in biomass analysis. Regression models (PLS and others) enable quantitative prediction of composition, often replacing lengthy chemical assays with a rapid spectral measurement (Acquah et al., 2016a). Classification models (PLS-DA, SVM, etc.) allow the sorting of biomass by type or quality, which can aid in feedstock verification and quality control (Jesus et al., 2024). Feature extraction methods like PCA help manage the complexity of spectral data and often improve model performance. The combination of FTIR and ML thus forms a powerful toolkit: FTIR provides a quick and rich measurement of the sample, and ML algorithms interpret that data to yield meaningful information. The success of this approach in the literature sets the stage for the present research, while also revealing certain gaps that the current study aims to fill.

2.3 Research Gap and Contribution

The review of existing literature indicates that while FTIR combined with machine learning is a promising strategy for biomass characterization, there are several limitations and open challenges in the current methodologies. Addressing these gaps is essential to further improve the accuracy, robustness, and utility of FTIR-based

analysis. This section identifies key research gaps and outlines how the present study will contribute to advancing the field by overcoming some of these limitations.

2.3.1 Limitations of existing approaches

A notable limitation in many studies to date is the reliance on linear multivariate techniques (like PLS regression or linear discriminant analysis) as the primary modeling tools. Linear models have the advantage of simplicity and often work well for roughly linear systems, but real spectral–composition relationships can be nonlinear or more complex than such models can capture. As an example, PCA, a linear dimensionality reduction method, treats variance in a straight-line manner and may fail to capture subtler, non-linear correlations in spectral data – especially when the dataset is limited in size (Jabed et al., 2023). PLS regression, while powerful, can sometimes overlook certain correlations or biases in the data; it assumes a linear relationship between latent factors and the response and can be sensitive to how data are scaled (Jabed et al., 2023). In practice, this means PLS might not fully account for interactions between spectral bands (for instance, when one constituent’s absorbance affects another’s baseline) or for nonlinear effects (such as saturation of absorbance at high concentrations). Consequently, the predictions for some components can suffer. Indeed, literature reports show differential success in predicting various components: some studies found excellent accuracy for lignin or extractives using PLS, but poorer accuracy for cellulose or hemicellulose (Acquah et al., 2016a). In Acquah et al. (2016), the PLS model’s performance for carbohydrates was weaker than for lignin (Acquah et al., 2016a), likely because cellulosic bands overlap more and have less distinct spectral features compared to lignin’s aromatic peaks (Javier-Astete et al., 2021). This points to a gap in capturing the full complexity of the spectra. Non-linear and interaction effects present in FTIR data (for example, peak shifts or broadenings at different concentrations) are not explicitly handled by linear methods. Therefore, one research gap is the limited exploration of advanced or non-linear ML algorithms in this domain. While techniques such as neural networks and tree ensembles have shown potential, they have not been as widely adopted in biomass FTIR analysis as one might expect, possibly due to historical reliance on chemometric tools. There is room to investigate whether employing these advanced models can significantly boost prediction accuracy across all components, especially for those (like holocellulose) that proved challenging for linear models.

Another limitation is the generality and robustness of the models developed. Many prior studies built calibration models on relatively homogeneous sets of samples – e.g., one species of wood, or a set of samples from a single experimental batch. While those models can perform well within that specific domain, they may not generalize to other biomass types or broader variations. Biomass is inherently variable: different species, growth conditions, harvest times, and pretreatments can all influence its composition and the resulting FTIR spectra. A model trained on (for instance) poplar wood might not directly apply to straw or grass, because the spectra could differ in baseline or specific band ratios (due to different lignin composition, mineral content, etc.). The transferability of models is a challenge – it often requires either recalibration or domain adaptation. In the literature, this gap is evident in that each study tends to develop a bespoke model for its own dataset, without demonstrating how it might be extended to others. Recently, some efforts have been made toward multi-feedstock models (calibrations that include multiple species or biomass types). Pushpa et al. (2024) is one such example, where a single model was developed for mixed feedstocks (Pushpa et al., 2024). Their success with an ANN on diverse biomass indicates it's feasible to create more universal models. Nonetheless, the general issue remains that we lack widely applicable models – each new feedstock often requires a new calibration. The present research sees an opportunity here: by incorporating a diverse training set (multiple biomass sources, broader property ranges) and using algorithms adept at handling variability, one can aim for a model that maintains accuracy across a spectrum of biomass types. This would significantly enhance the practical utility of FTIR-ML methods (e.g., in industry, one model could potentially handle various feedstocks encountered, rather than maintaining separate models for each). The gap in model robustness also ties into how models are validated; some prior works did not rigorously test model performance on independent sample sets (external validation), leaving uncertainty about how they perform on truly unseen data.

Data scarcity is another concern. Building any data-driven model is constrained by the availability of quality training data (here, samples with known composition and corresponding spectra). Preparing such datasets is resource-intensive, since each sample's composition must typically be measured by the reference chemical methods to serve as ground truth. As a result, many studies operate with limited sample sizes (sometimes only on the order of tens of samples for calibration). This can limit the

complexity of the model that can be reliably trained and increase the risk of overfitting specific spectral quirks of the training set. The gap here is not just in quantity of data, but in consistency and coverage of data – ensuring the calibration covers the range of compositions and sample types expected in application. Some researchers have highlighted the need for more robust validation approaches in this context. Fadlelmoula et al. (2023), in a review of FTIR-ML for biological samples, emphasized that multiple ML approaches should be compared and rigorous criteria used for model selection and validation (Fadlelmoula et al., 2023). Although their focus was biomedical, the principle applies to biomass: to truly advance the field, studies must adhere to high standards of model assessment (such as using separate test sets, reporting figures of merit like RMSEP, R^2 , RPD, etc., and avoiding overfitting). Without such standards, it is hard to identify the best methods or to combine insights across studies. This thesis recognizes that gap in methodology rigor and aims to implement best practices in model development (e.g., using cross-validation and external validation, and statistically comparing different modeling techniques on the same dataset) to provide more reliable conclusions.

Perhaps one of the most interesting gaps is in the interpretation of ML models and spectra. Much of the existing work treats the ML model as a means to an end (predicting composition accurately), but gives less attention to what the model reveals about the spectral features themselves. In other words, the models can be black boxes – they predict lignin content, but we might not know which wavelengths were most influential in that prediction. From a scientific standpoint, interpreting the model can yield valuable information: it could confirm known correlations (e.g., that the 1510 cm^{-1} band is indeed a major contributor to lignin predictions, which aligns with chemical knowledge) or even discover new ones (e.g., maybe a combination of absorbances at unexpected regions correlates with a property, pointing to a previously unnoticed marker). Some recent studies outside the narrow realm of biomass have started to report feature importance and chemical interpretation of models (Jabed et al., 2023), reflecting a growing awareness that explainable AI techniques can and should be applied in spectroscopy. In biomass analysis, however, this approach is not yet commonplace. The gap, therefore, is that we lack a deep understanding of how exactly ML models are leveraging the FTIR data. Bridging this gap could enhance trust in these models (which is important for industry adoption) and ensure that the

predictions make chemical sense. For instance, if a model were to erroneously rely on a noise spike or an artifact, interpretability checks might catch that issue. Conversely, if a model highlights an unexpected spectral region as important, analysts can investigate that region for potential chemical reasons (perhaps indicating the presence of a minor compound or some interference). The current literature seldom discusses such interpretation; they report accuracy metrics but do not always link them back to spectral features. This is an area the present research will address by incorporating interpretability as a core component of the analysis.

Contributions of the Present Study: In light of the above gaps, this thesis aims to push the boundaries of FTIR-based biomass analysis with machine learning in several ways:

2.3.2 Adoption of advanced ML algorithms for improved accuracy

We will go beyond the standard PLS approach and evaluate more complex models (such as support vector machines, random forest ensembles, and neural networks) on the same dataset to determine if improvements in predictive accuracy can be achieved. By doing so, we address the gap regarding linear vs. non-linear modeling. For example, if PLS regression plateaus in performance for predicting cellulose content, we will test whether an ANN can capture additional non-linear patterns to reduce prediction error (Pushpa et al., 2024). Similarly, we will explore ensemble techniques; if prior knowledge suggests that certain spectral regions are especially informative for a given component, a tree-based model might naturally leverage that by splitting on those features. A comparative approach will be taken, where models are trained and tested under identical conditions (using rigorous cross-validation and external test sets) so that we can quantitatively assess the gains. The expectation is that at least for some constituents (especially those with more complex spectral signatures or lower concentrations), advanced ML will yield higher accuracy and lower uncertainty in predictions than the classical methods. Achieving a measurable improvement in predictive performance (e.g., higher R^2 and RPD, lower RMSE) would be a significant contribution, as it would demonstrate a path forward for more reliable biomass analysis. It would also corroborate the indications from recent studies that, for multi-component systems, embracing non-linearity (through ML) can pay off in better models (Pushpa et al., 2024). Improving accuracy has practical implications: for

instance, more precise knowledge of composition can lead to better control in bioprocessing or biomass valuation.

2.3.3 Enhanced model generality and robustness

This study is designed with an emphasis on broad applicability. We plan to incorporate a diverse array of biomass samples – potentially including different species (hardwood, softwood, agricultural residues) and samples with a wide range of compositions – to train the models. By doing so, we directly tackle the gap of narrow model scope. A successful outcome would be a single model (or a set of models with minimal tailoring) that maintains good accuracy across this diversity. Such a model could be truly useful in scenarios where feedstock may vary. In the process of building this, we will also explore data augmentation or calibration transfer techniques (if needed) to handle any systematic differences, say, between spectra of different particle sizes or slight instrument variations. The contribution here would be demonstrating a more universal calibration – essentially extending the work of previous researchers who each looked at one feedstock, by merging and learning from multiple feedstocks in one unified approach. Additionally, all models will be subjected to robust validation: for example, using one biomass type as an independent test while training on others, to see how well the model extrapolates. If we find the model struggling with a particular type, that will be instructive and could lead to methodologies like domain-specific sub-models or inclusion of dummy variables to indicate biomass class within the model. Either way, documenting the journey towards a general model will provide guidance for future work. By improving robustness, we intend to reduce the need for recalibration – a step towards practical deployment of FTIR-ML methods outside of controlled lab settings.

2.3.4 Interpretability and spectral insight

A distinguishing feature of this research is the strong focus on interpreting the ML models in chemical terms. Rather than treating the models as opaque prediction engines, we will apply techniques to open the black box. For PLS models, this is naturally done by examining the loading weights (which wavelengths contribute most to each latent factor) and the regression coefficients (which approximate the net effect of each original wavelength on the prediction). For non-linear models, we will use methods like variable importance (available in RF models) and sensitivity analysis (for ANNs, seeing how slight changes in input at a given wavelength affect the output).

We may also employ modern interpretation tools such as SHAP (SHapley Additive exPlanations) values to consistently rank the influence of each spectral region on the predictions. By correlating these findings with known band assignments, we expect to validate that the models are grounding their decisions in sensible spectral features. For example, if the model identifies a region around 1230 cm^{-1} as important for holocellulose prediction, we can relate that to C–O stretching in lignin and cellulose (which would make sense) (Javier-Astete et al., 2021). In the event an important model feature does not correspond to a known band, that will be investigated – it could indicate an artifact or perhaps a previously under-recognized marker (such as a combination band or minor component signal). By reporting such details, the study contributes a layer of insight often missing in prior work. This not only enhances scientific understanding but also builds confidence for users: an industry practitioner would be more inclined to trust a predictive model if told, for instance, that “the model bases its lignin prediction largely on the aromatic absorbance at 1510 cm^{-1} and associated overtone features,” which aligns with chemical expectations, rather than the model being a mysterious mathematical construct. Recent publications in related fields have stressed the value of reporting feature significance and ensuring the ML model’s behavior can be chemically interpreted (Jabed et al., 2023), and this work will explicitly follow that ethos.

2.3.5 Integrative improvement and practicality

Finally, by addressing accuracy, generality, and interpretability together, this study aims to deliver an integrated improvement to the state-of-the-art. The end goal is a machine learning-enhanced FTIR analysis approach for biomass that is more accurate, more generalizable, and more transparent than those previously reported. This can be seen as moving the technology closer to real-world application. For example, a robust and interpretable model could be implemented for routine biomass feedstock analysis in a biorefinery, providing near real-time data on composition (moisture, cellulose, lignin, etc.) without lengthy lab tests. The benefits echo what has been observed in other domains using FTIR-ML – the method can be simple, rapid, and cost-effective, requiring neither chemical reagents nor expensive analytical equipment beyond the spectrometer (Fadlelmoula et al., 2023). By demonstrating improvements, this research provides a template for such practical deployment. Furthermore, any novel insights (such as an understanding of which spectral regions carry predictive power

and which do not) add to the foundational knowledge that future researchers can use in method development.

In summary, the literature reveals that FTIR combined with machine learning is a powerful approach for biomass analysis, but also that current implementations have room for enhancement in accuracy, scope, and clarity. The contribution of this thesis lies in systematically pushing those fronts: using a range of ML models to seek better performance, constructing models on broader datasets for wider applicability, and embedding interpretability into the modeling process. By doing so, it advances the field toward a more accurate, general, and interpretable use of spectroscopic data for biomass characterization. These advancements aim not only to fill the gaps identified in academic research but also to pave the way for real-world analytical solutions in biomass utilization industries, where rapid and reliable composition analysis is critically needed. Ultimately, the study endeavors to show that with modern ML techniques and careful methodology, FTIR-based biomass analysis can achieve higher precision and insight, strengthening its role as a key tool in bioenergy and bioproduct research. The following chapters will detail the materials and methods used to realize these objectives, and present the results that support these contributions (Javier-Astete et al., 2021).

3. MATERIALS AND METHODS

3.1 Biomass Sample Collection and Preparation

This study utilized a diverse set of 56 biomass samples representing various lignocellulosic materials. The samples included agricultural residues (e.g. barley meal, rice straw, sunflower stalks), nut shells and hulls (almond shell, walnut shell), woody biomass (walnut branch, pine cone, ash bark), and other plant by-products (pea stalk, grape seed, tea waste, etc.). Each sample was first air-dried or oven-dried to reduce moisture content to low levels (often <10%) (Hames et al., 2008). Oven drying was performed at a moderate temperature (around 45–105 °C) until reaching a constant weight, ensuring removal of free moisture without chemically altering the biomass (Hames et al., 2008). After drying, the biomass was milled and sieved to a fine powder (typically to <250 micrometer particle size) to obtain homogeneous samples for analysis. Fine grinding of the samples is important because FTIR measurements and calibration models require dry, uniformly ground biomass for consistent and accurate spectra (Whatley et al., 2023).

After preprocessing, each biomass was characterized by both “structural analysis” and “proximate analysis” to quantify its composition. Structural analysis refers to the main lignocellulosic components: extractives, holocellulose (combined cellulose and hemicellulose), and lignin. Extractive content was determined by solvent extraction (e.g. successive Soxhlet extractions with organic solvents), which removes non-structural compounds like fats, resins, and phenolics. The purpose of this step is to eliminate substances that could interfere with subsequent analysis of structural polysaccharides and lignin. Holocellulose (the total polysaccharide fraction) was obtained either by summing the cellulose and hemicellulose content or by a direct method (such as sodium chlorite delignification) that removes lignin and leaves a holocellulose residue (Javier-Astete et al., 2021). Lignin content was measured as the residue remaining after strong acid hydrolysis of the biomass (Klason lignin method), following standardized protocols (NREL, n.d.). For completeness, these structural

components were often reported on a dry, ash-free basis to facilitate comparison across samples (i.e. normalized to remove moisture and inorganic content). Proximate analysis parameters: moisture (inherent water content), volatile matter, ash, and fixed carbon. Moisture was measured by drying a sample at 105 °C and noting the weight loss (NREL, n.d.). Ash content was determined by igniting the sample in a muffle furnace at ~575 °C to 600 °C until all organic matter was combusted, leaving only mineral residue (NREL, n.d.). Volatile matter was determined by heating the sample to 950 °C under an inert atmosphere and measuring the weight loss excluding moisture, and fixed carbon was computed by difference (100% – moisture – ash – volatile). These analyses yielded values such as: extractives ~5–37%, holocellulose ~43–91%, lignin ~1.5–40% (varying widely by sample type), moisture ~4–9%, ash ~0.8–16%, etc., reflecting the broad range of biomass compositions in the dataset. All analytical procedures followed standard methods in biomass analysis (NREL, n.d.), ensuring that the input data (composition percentages) were accurate and comparable. The resulting dataset thus contained, for each of the 56 samples, a profile of its chemical composition (extractives, holocellulose, lignin, moisture, volatile matter, ash, fixed carbon), which served as the input features for modelling, as well as its corresponding FTIR spectral data as described below.

3.2 FTIR Spectroscopy

FTIR measurements were carried out in the wavenumber range of 4000–450 cm^{-1} using the Perkin Elmer Spectrum Two spectrometer, which captures the main vibrational bands associated with lignocellulosic biomass components. Measurements were performed using the ATR mode without replicate scans, as replicate measurements are not standard practice for this device. After each measurement, the ATR crystal and the upper contact surface were wiped clean with a paper towel to avoid cross-contamination. In cases where residues remained on the surface, ethanol was used for cleaning; however, such a need did not arise for the biomass samples in this study. Under these conditions, typical FTIR spectra exhibited broad O–H stretching bands (around 3400 cm^{-1}), C–H stretching (near 2920 cm^{-1}), and a series of peaks in the “fingerprint” region (1800–800 cm^{-1}) corresponding to functional groups of cellulose, hemicellulose, and lignin. The resulting spectral data consisted of

absorbance (or transmittance) values at 3551 discrete wavenumbers, providing a high-dimensional chemical signature for each biomass sample.

To ensure the spectral data were reliable for subsequent analysis, several preprocessing steps were applied to the raw FTIR spectra. Baseline correction was performed to remove any sloping or offset of the spectrum baseline, which can occur due to scattering by particles or ATR crystal imperfections. This was done by fitting a baseline (using polynomial or rubber-band algorithms) and subtracting it, so that the absorbance baseline around non-absorbing regions (e.g. $\sim 3800\text{--}4000\text{ cm}^{-1}$) was near zero (Tkachenko & Niedzielski, 2022). Next, each spectrum was normalized to account for differences in sample quantity or path length. In practice, a simple normalization (such as unit vector normalization or setting a reference peak to a constant value) was used so that all spectra are on a comparable scale (Tkachenko & Niedzielski, 2022). This ensures that variations in spectral intensity reflect true compositional differences rather than sample concentration. Additionally, noise filtering was applied to improve spectral quality. A Savitzky–Golay smoothing filter (or similar moving average technique) was used to reduce high-frequency noise while preserving peak shapes (Shimadzu, n.d.). This slight smoothing makes it easier to detect genuine peaks, at the cost of a very minor reduction in resolution. In some cases, spectral derivatives (e.g. 1st or 2nd derivative spectra) were examined as well, since taking derivatives can help resolve overlapping peaks and correct baseline shifts (Tkachenko & Niedzielski, 2022). However, for the main analysis we retained the processed zero-order spectra after baseline correction, normalization, and smoothing. The final prepared spectral dataset was a matrix of dimension 56 samples $\times$ 3551 wavenumbers, with each row representing a pre-processed FTIR spectrum of a biomass sample. These spectra encapsulate the chemical fingerprint of each sample and serve as the target outputs in our modelling approach.

3.3 Machine Learning Approach

This section describes how we constructed the dataset, selected machine learning models, performed feature engineering, and organized training/validation procedures for Phase 1, Phase 2, and Phase 3 of the thesis. The overall aim is to predict features of the FTIR spectrum (from entire spectra to broad or narrow spectral dips) using nine key biomass characteristics as inputs.

3.3.1 Dataset construction

Each biomass sample in this study is characterized using nine input features that capture both compositional and categorical properties. The first feature is the biomass category, a categorical label that classifies the sample into one of several types such as woody biomass, herbaceous biomass, or other relevant groups. This category, comprising seven distinct classes, is transformed into a machine-readable format using techniques like one-hot encoding to ensure compatibility with machine learning algorithms. The remaining eight features are numerical variables, expressed predominantly as percentages, and represent key compositional metrics of the biomass. These include humidity, volatile matter content, ash percentage, and fixed carbon. A derived metric, computed as 100 minus the sum of moisture and ash, serves as a normalized indicator of the organic fraction of the sample. The last three variables denote the dry ashless percentages of extractive substances, holocellulose, and lignin—factors that provide chemically refined insights by removing the influence of water and inorganic content.

Collectively, these nine features form the predictor set for all machine learning models developed in this thesis. They were selected to encapsulate the fundamental chemical and physical characteristics of each biomass sample, ensuring that the models are equipped with the necessary information to make meaningful predictions regarding the spectral response.

The nature of the model output varies depending on the modeling phase. In Phase 1, which is focused on full-spectrum regression, the output consists of the complete FTIR transmittance spectrum for each sample. This spectrum spans thousands of discrete wavenumber points across the mid-infrared region, creating a high-dimensional, multi-output regression problem. The objective in this phase is to predict the transmittance—or alternatively, absorbance—intensity at each wavenumber using only the nine input variables. The predicted spectra reflect underlying molecular vibrations and chemical functionalities, offering insight into the sample's structural composition. The experimental FTIR transmittance spectra demonstrate characteristic absorption patterns corresponding to distinct functional groups, which are essential for interpreting the chemical makeup of lignocellulosic biomass.

In Phase 2, the modeling task shifts from predicting detailed spectral intensities to classifying whether absorbance peaks appear within broader wavenumber intervals. Eight such intervals are defined: 4000–3700 cm⁻¹, 3700–3000 cm⁻¹, 3000–2800 cm⁻¹, 2800–1800 cm⁻¹, 1800–1500 cm⁻¹, 1500–1150 cm⁻¹, 1150–900 cm⁻¹, and 900–450 cm⁻¹. For each interval, a binary classification is performed, assigning a value of “1” if a pronounced absorbance peak (or corresponding transmittance dip) is present in the spectral region, and “0” otherwise. This transformation of the output into a set of eight binary values per sample frames the problem as a multi-label classification task.

These spectral regions correspond to major chemical bond vibrations and are associated with specific functional groups. For instance, the 4000–3000 cm⁻¹ range encompasses broad O–H and N–H stretching modes, while the 3000–2800 cm⁻¹ range is dominated by C–H stretching vibrations characteristic of aliphatic structures. The interval from 1800 to 1500 cm⁻¹ captures carbonyl (C=O) and aromatic C=C absorptions, which are typical of lignin and certain hemicellulose components. The subsequent intervals cover the fingerprint region of the spectrum, rich in polysaccharide and aromatic signals crucial for distinguishing holocellulose from lignin content. Table 3.1 summarizes these wavenumber intervals along with their corresponding functional group assignments, providing a foundation for interpreting the chemical relevance of each region in relation to biomass composition.

Table 3.1 : FTIR spectra intervals.

Wavenumber Range (cm ⁻¹)	Bond Type / Vibration	Functional Groups
4000-3700	O–H stretching (free – OH)	“Free” O–H groups (non-hydrogen-bonded hydroxyl – e.g. isolated alcohol or phenol – exhibit sharp O–H stretch in ~3600–3700 cm ⁻¹ range) (Dai et al., 2023)
3700-3000	X–H stretching (light atoms): includes =C–H, ≡C–H, O–H, N–H stretches	Alkenes & aromatics (≡C–H stretch ~3020–3100 cm ⁻¹) ; terminal alkynes (≡C–H stretch ~3300 cm ⁻¹ , typically a sharp peak); alcohols & phenols (O–H stretch ~3200–3600 cm ⁻¹ , broad if hydrogen-bonded, sharp if “free”); carboxylic acids (O–H extremely broad, often centered ~3000 cm ⁻¹); 1° and 2° amines/amides (N–H stretches ~3300–3500 cm ⁻¹ ; primary –NH ₂ shows dual peaks) (IR Absorption Frequencies, 2014)
3000-2800	C–H stretching (sp ³ C–H); formyl C–H stretching (aldehyde)	Alkanes (sp ³ C–H stretches ~2850–2960 cm ⁻¹); aldehydes (C–H stretch of –CHO appears as two weak bands ≈ 2900 and 2720 cm ⁻¹ due to Fermi resonance) (IR Absorption Frequencies, 2014)

Table 3.1 (continued) : FTIR spectra intervals.

Wavenumber Range (cm ⁻¹)	Bond Type / Vibration	Functional Groups
2800-1800	C≡C stretching (alkyne); C≡N stretching (nitrile); S–H stretching (thiol); broad O–H (carboxylic acid); aldehydic C–H stretch (Fermi band)	Alkynes (C≡C stretch ~2100–2260 cm ⁻¹ , typically weak); nitriles (C≡N stretch ~2260–2240 cm ⁻¹); thiols (S–H stretch ~2550 cm ⁻¹); carboxylic acids (O–H broad absorption often spanning ~2500–3300 cm ⁻¹ , centered ~3000); aldehydes (formyl C–H weak “finger” peaks near ~2850 and 2750 cm ⁻¹) (Andrade et al., 2008; IR Absorption Frequencies, 2014)
1800-1500	C=O stretching (carbonyl); C=C stretching (alkene & aromatic); N–O asymmetric stretching (–NO ₂)	Carbonyl groups – e.g. ketones, aldehydes, esters, carboxylic acids, amides (strong C=O absorbances typically ~1650–1750 cm ⁻¹ , depending on conjugation and type); unsaturated C=C bonds – alkenes (~1640–1680 cm ⁻¹) and aromatic rings (~1600 cm ⁻¹); nitro compounds (NO ₂ asymmetric stretch ~1530–1550 cm ⁻¹) (IR Absorption Frequencies, 2014; Joseph M. Fox, 2013)
1500-1150	C–H bending (deformation of CH ₂ /CH ₃); C=C stretching (aromatic ring); N–O symmetric stretching (–NO ₂)	Alkanes (methyl and methylene C–H bends at ~1465, 1450, 1375 cm ⁻¹); aromatic compounds (ring C=C stretches ~1500–1600 to 1400 cm ⁻¹); nitro compounds (NO ₂ symmetric stretch ~1350 cm ⁻¹) (IR Absorption Frequencies, 2014)
1150-900	C–O single-bond stretching; C–N stretching (aliphatic); C–F stretching	Alcohols, phenols, ethers, esters (C–O stretch in 1000–1150 cm ⁻¹ range); amines (C–N stretch ~1020–1250 cm ⁻¹ for aliphatic amines); organofluorines (C–F ~1100 cm ⁻¹ , strong) (Alfred D. Bacher, 2016; IR Absorption Frequencies, 2014)
900-450	C–X stretching (X = Cl, Br, I); out-of-plane C–H bending (aromatic)	Alkyl halides (C–Cl ~800–600; C–Br ~600–500; C–I ~500 cm ⁻¹); aromatic rings (characteristic C–H bending patterns below ~900 cm ⁻¹) (Alfred D. Bacher, 2016; Joseph M. Fox, 2013)

In the third phase, the focus is narrowed to three specific wavenumber intervals that correspond to functionally important molecular vibrations. These intervals are: 3000–2800 cm⁻¹, which is typically associated with aliphatic C–H stretching vibrations; 1800–1500 cm⁻¹, covering the carbonyl and aromatic region; and 1150–900 cm⁻¹, known as the carbohydrate fingerprint region. For each of these intervals, a binary indicator is used to represent the presence or absence of a well-defined spectral peak. This results in three classification outputs, each corresponding to one of the targeted regions. The rationale behind this approach is to isolate and emphasize spectral regions that are functionally significant, such as those related to lignin or cellulose content.

Consequently, the dataset can be understood as being divided into three conceptual sub-datasets. The first supports full-spectrum regression, aiming to model continuous outcomes across the entire spectral range. The second facilitates broad-range classification, which focuses on more general spectral patterns. The third concentrates on narrow-range, targeted classification, specifically within the predefined critical intervals. Although all three sub-datasets utilize the same set of nine input variables,

they differ in the nature of their response outputs and the objectives of the predictive models applied.

3.3.2 Machine learning model selection

The first phase involves full-spectrum regression, which is framed as a multi-output regression problem. In this context, each individual wavenumber in the spectral range is treated as a separate output variable, potentially resulting in thousands of simultaneous predictions. This high dimensionality, along with the collinearity inherent in spectral data, necessitates the use of specialized regression techniques. PLSR and Ridge Regression are commonly applied due to their robustness in handling multicollinearity and high-dimensional feature spaces. Although more complex models such as Random Forest Regressors or Neural Networks can be applied in a multi-output setting, the relatively small sample size in this study (56 samples) favors simpler models with regularization, which tend to generalize more reliably under such constraints.

The second phase transitions from regression to broad-range classification. In this formulation, each sample is associated with eight binary classification targets, each corresponding to a broad spectral interval. Two modeling strategies are considered: training separate binary classifiers for each interval, or employing a multi-label classification approach where a single model outputs all eight binary labels simultaneously. Suitable algorithms for this task include Logistic Regression, Random Forest, Gradient Boosted Trees, and SVM with carefully selected kernels and regularization parameters. Compared to full-spectrum regression, this classification task is generally less complex, as the model is required only to determine the presence or absence of a prominent absorption dip in each defined spectral region.

The third phase further refines the classification task by focusing on narrow spectral intervals of particular chemical relevance. Specifically, attention is directed toward the wavenumber ranges 3000–2800 cm^{-1} (aliphatic C–H stretching), 1800–1500 cm^{-1} (carbonyl and aromatic region), and 1150–900 cm^{-1} (the carbohydrate fingerprint region). Each interval produces a single binary output indicating the presence or absence of a distinct spectral peak. This targeted approach is structurally similar to the broad-range classification in Phase 2 but benefits from increased specificity, as the selected intervals are more directly associated with chemically meaningful

components such as lignin or holocellulose. As with the previous phase, classification models such as logistic regression, SVMs, and tree-based ensembles are employed. However, due to the narrower spectral focus and stronger chemical signal-to-noise characteristics, these models often yield higher classification accuracy.

Across all three phases, model development is supported by systematic hyperparameter tuning and algorithm comparison. This process is essential in identifying the most effective methods for linking the nine input variables to their respective spectral outcomes, thereby ensuring both interpretability and predictive performance.

3.3.3 Feature engineering & data preprocessing

Prior to model training, several preprocessing steps are applied to prepare the dataset for machine learning. These steps ensure that the input features are appropriately scaled, encoded, and free from inconsistencies that could negatively affect model performance.

First, data normalization is performed on the eight numeric composition features, such as moisture and ash content. Since these features can vary significantly in magnitude, they are standardized to have zero mean and unit variance. This normalization step is particularly important for learning algorithms that are sensitive to feature scales, such as distance-based models or regularized linear models.

Next, the biomass category feature, which is a categorical variable with seven distinct classes, is encoded to enable its use in models that require numerical input. Depending on the model type, this feature is either one-hot encoded—resulting in a binary vector for each category—or treated as an integer-coded label. The choice of encoding strategy is made with respect to model compatibility and performance considerations.

To address missing data, imputation techniques are applied. If any of the composition or categorical fields contain missing values, these are replaced using either mean imputation (for continuous variables) or k-nearest neighbors (KNN) imputation, ensuring that the final dataset contains no null entries. This step is crucial for maintaining model robustness and avoiding errors during training and evaluation.

Although the dataset includes only nine input features, dimensionality reduction is still considered due to potential redundancy among features. For example, some features

are algebraically related, such as fixed carbon or calculated values like "100 – moisture – ash". Multicollinearity is evaluated using correlation analysis and variance inflation factors. Where appropriate, highly collinear features may be removed or combined to reduce redundancy and improve model interpretability.

The structure of the output data is customized according to the specific objectives of each modeling phase. In Phase 1, the entire FTIR spectrum serves as a high-dimensional target vector for regression. In Phase 2, the spectrum is partitioned into six broad intervals, and binary labels are assigned to indicate the presence or absence of dominant spectral features in each region. In Phase 3, the focus is further refined to three narrow intervals of particular chemical relevance, each labeled to reflect the presence or absence of a well-defined peak.

By organizing the data so that each sample is represented by a consistent set of nine input features—including the encoded biomass category—and the appropriate output format depending on the modeling phase, a unified and scalable preprocessing pipeline is established. This consistency facilitates seamless training, validation, and evaluation across all modeling tasks.

3.3.4 Training & validation

The dataset used in this study comprises a total of 56 samples. Given the limited sample size, careful data partitioning is essential to ensure reliable model evaluation. To this end, an 80/20 split is employed, resulting in approximately 45 samples allocated for training and 11 reserved for testing. This split strikes a balance between maximizing training data availability and maintaining a representative test set for final model evaluation.

Within the training set, model selection and hyperparameter tuning are performed using k-fold cross-validation, typically with $k = 5$. This method divides the training data into five subsets, using four for model training and one for validation in each iteration. By cycling through all possible folds, this approach helps mitigate overfitting and provides a more stable estimate of model performance, especially in the context of small datasets. Hyperparameters tuned during this process include the number of components in PLS regression, maximum tree depth in ensemble models, and regularization parameters such as C and γ in SVM.

Evaluation metrics are selected according to the modeling objective of each phase. For Phase 1, which involves regression over the full FTIR spectrum, performance is assessed using RMSE and the R^2 , which respectively quantify the prediction error magnitude and the proportion of variance explained by the model. In contrast, Phases 2 and 3 involve classification tasks. In these phases, key evaluation metrics include overall accuracy, micro-averaged F1 scores, and confusion matrices, which collectively assess the model's ability to correctly classify the presence or absence of a spectral dip in each region. For multi-label classification settings, both individual label performance and aggregate metrics are reported to provide a comprehensive evaluation.

Following hyperparameter optimization, the final model is retrained on the entire training dataset and then evaluated on the held-out test set. This ensures that the final performance metrics reflect the generalization ability of the model to unseen data.

By maintaining a consistent set of nine input features and applying a unified, rigorous training protocol across all phases, it becomes possible to systematically assess how well biomass composition predicts varying levels of spectral detail. This phased approach allows for a direct comparison between models targeting full-spectrum regression, broad-range classification, and narrow-range classification, thereby providing insights into the granularity of spectral information that can be inferred from compositional data.

4. RESULTS AND DISCUSSION

4.1 Model Performance (Phase 1: Full-Spectrum Regression)

In Phase 1, we treated the prediction of every wavenumber's intensity in the FTIR spectrum as a multi-output regression problem, aiming to see if the nine biomass composition features could reconstruct the entire infrared profile. Below, we present the results from the four regression models evaluated—PLS, Ridge Regression, Random Forest, and a MLP neural network.

4.1.1 Regression metrics and comparisons

As illustrated in Figure 4.1, several regression models were evaluated for their ability to predict the full FTIR spectrum from the nine input features. Among the models tested, PLS regression achieved the lowest RMSE, with a value of 1.031. This indicates a moderate degree of deviation between predicted and actual spectral intensities and reflects PLS's strength in managing multicollinearity within high-dimensional spectral data.

Ridge Regression, which also incorporates regularization to handle correlated inputs, produced a slightly higher RMSE of 1.106. This performance suggests a somewhat reduced predictive accuracy relative to PLS, though still within a reasonable range.

The Random Forest model yielded an RMSE of 1.055, which is broadly comparable to the PLS result but not as low. This outcome indicates that non-linear models may provide competitive performance in this setting, though not necessarily superior without further tuning.

Overall, the results presented in Figure 4.1 highlight PLS as the most effective regression approach among those tested, given the dataset's characteristics and the full-spectrum prediction objective.

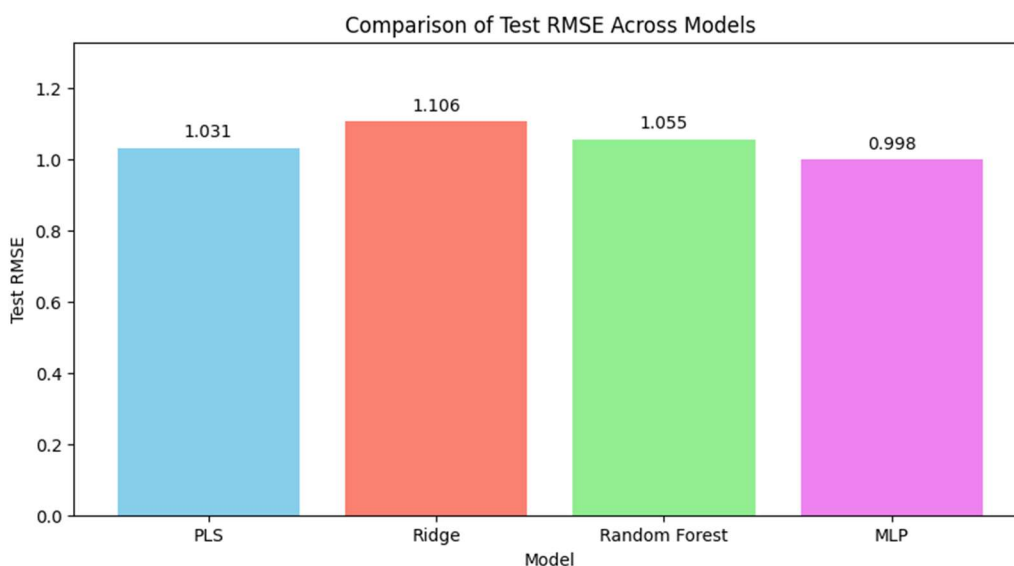


Figure 4.1 : Comparison of test RMSE across models.

As shown in Figure 4.2, the coefficient of determination (R^2) was used to evaluate how effectively each regression model captured variance in the full FTIR spectral data. PLS regression achieved an R^2 value of 0.168, indicating that it was able to explain approximately 17% of the total variance. While modest, this result suggests that PLS can extract some meaningful structure from the compositional features despite the complexity and high dimensionality of the output space.

Ridge Regression, by contrast, performed less favorably, with an R^2 of just 0.043. This implies that its predictions accounted for only about 4% of the spectral variance, highlighting its relative weakness in capturing the intricate multi-output relationships inherent in this task.

The Random Forest model demonstrated a slight improvement over Ridge Regression, achieving an R^2 of 0.130. While this value remains lower than that of PLS, it suggests that tree-based methods can offer reasonable performance, particularly when non-linear relationships are present in the data.

Notably, the MLP model yielded the highest R^2 value among the models tested, at 0.210. Although still limited in absolute terms, this result indicates that the MLP captured over 21% of the spectral variance, outperforming all other approaches in this comparison.

Together, the results depicted in Figure 4.2 emphasize the challenges posed by full-spectrum prediction using a small dataset and underscore the potential of more

flexible, non-linear models—such as MLPs—in capturing complex spectral patterns when data quantity and quality permit.

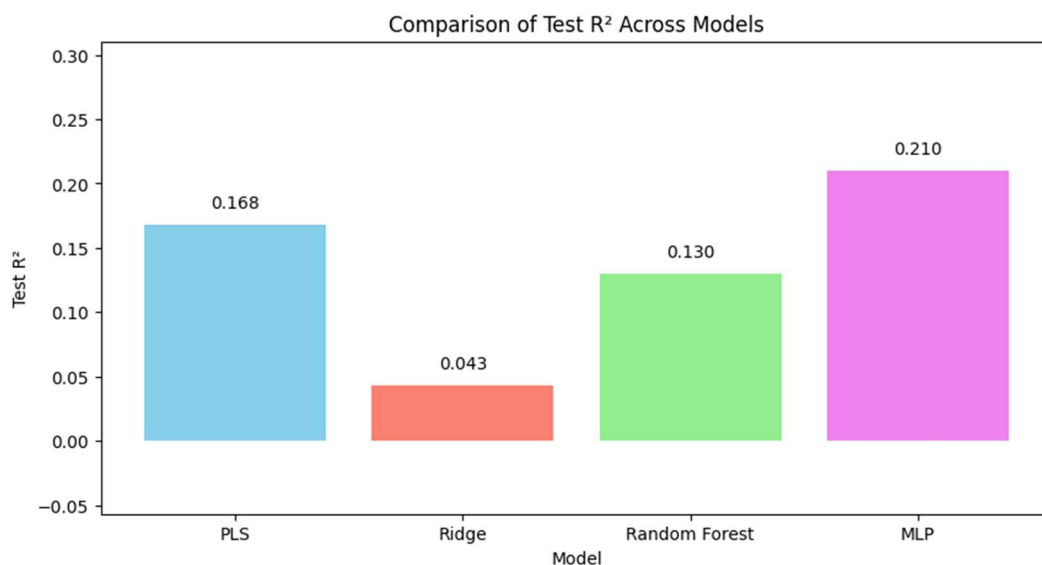


Figure 4.2 : Comparison of test R^2 across models.

Overall, MLP exhibited the most promising performance on both error (lowest RMSE) and explained variance (R^2 around 0.21). The results also highlight the high complexity of this full-spectrum prediction task, given the modest R^2 values across all methods—an expected outcome given the limited sample size and the thousands of possible outputs.

4.1.2 Visualizing predicted vs. actual spectra

Figure 4.3 presents a visual comparison between the true FTIR spectrum of a representative test sample (depicted by the solid blue line) and the predicted spectra generated by each of the regression models (shown as colored dashed lines). Several key observations emerge from this comparison.

First, there is a noticeable intensity offset in many of the model predictions. In particular, PLS, Ridge Regression, and Random Forest frequently overestimate the transmittance values in certain spectral regions, resulting in predicted curves that consistently lie above the actual spectrum. This effect is especially pronounced in the $900\text{--}1800\text{ cm}^{-1}$ range, suggesting a systematic bias in these models when approximating moderate absorption features.

The MLP model, represented by the magenta dashed line, demonstrates a comparatively closer alignment with the true spectrum across multiple regions. This

improved visual correspondence is consistent with its lower RMSE, as discussed previously, and reflects the MLP’s capacity to model more complex, non-linear relationships between the input features and the spectral output.

Nonetheless, accurately capturing sharp dips in transmittance—such as those near 1650 cm^{-1} and within the $2800\text{--}3000\text{ cm}^{-1}$ range—remains a challenge for all models. These deep absorption features are difficult to predict given the limited input space (only nine compositional variables), highlighting the inherent complexity of the full-spectrum prediction task.

Overall, the visualizations in Figure 4.3 reinforce the quantitative findings: while no model is able to replicate the FTIR spectrum with complete fidelity, the MLP, and to a lesser extent PLS, exhibit a closer approximation of the spectral shape, particularly in regions with gradual intensity changes.

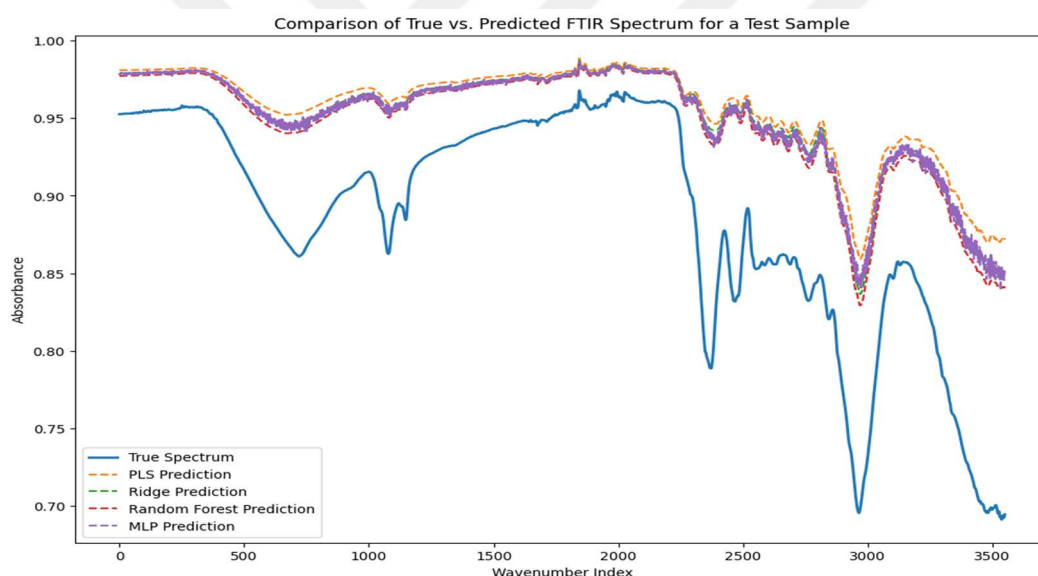


Figure 4.3 : Comparison of true vs predicted FTIR spectrum for a test sample.

4.1.3 Interpretation and discussion

The task of predicting thousands of FTIR wavenumber intensities from only nine compositional features and a total of 56 samples represents a classic high-dimensional regression problem. This setting presents significant challenges due to both collinearity among spectral outputs and the limited number of training observations. Many of the wavenumber intensities are strongly correlated with one another, making methods such as PLS regression particularly well-suited. This explains PLS’s relatively competitive performance, as it is specifically designed to extract latent

variables that capture shared structure between inputs and outputs. In contrast, Ridge Regression—while incorporating regularization to manage multicollinearity—may struggle to fully capture the non-linear relationships embedded in the data, which contributes to its comparatively lower performance.

Non-linear models such as Random Forest and MLP offer a distinct advantage in modeling complex interactions between input features and spectral outputs. However, their generalization ability is constrained by the small sample size. Despite this limitation, the MLP model outperformed the other approaches, achieving the best results in terms of both RMSE and R^2 . This suggests that a carefully tuned neural network can uncover subtle, non-linear patterns that link biomass composition to spectral variation, even in low-data regimes.

While the results from Phase 1 provide an important baseline for modeling the full FTIR spectrum, the relatively low R^2 values across all models underscore the difficulty of predicting detailed spectral signatures directly from limited compositional information. This motivates the subsequent phases of the analysis—Phases 2 and 3—which explore whether reformulating the prediction task to focus on simplified outputs, such as broad-interval or narrow-peak classification, can yield improved predictive performance and more interpretable chemical insights.

In conclusion, Phase 1 demonstrates that full-spectrum regression is technically feasible but remains a demanding problem when working with a small and highly collinear dataset. Among the models tested, the MLP achieved the best overall performance, with an RMSE of approximately 0.998 and an R^2 of 0.210, followed by PLS, Random Forest, and Ridge Regression. These findings establish a performance benchmark and provide critical context for evaluating the simplified classification tasks addressed in Phases 2 and 3.

4.2 Model Performance (Phase 2: Broad-Range Classification)

In Phase 2, the prediction task is reformulated to move away from reproducing every individual wavenumber intensity across the FTIR spectrum, as in Phase 1, and instead focuses on identifying whether a pronounced transmittance dip—corresponding to an absorbance "peak"—is present within eight predefined broad spectral intervals. These intervals, chosen to span the full infrared range while maintaining chemical relevance,

are defined as follows: 4000–3700 cm^{-1} , 3700–3000 cm^{-1} , 3000–2800 cm^{-1} , 2800–1800 cm^{-1} , 1800–1500 cm^{-1} , 1500–1150 cm^{-1} , 1150–900 cm^{-1} , and 900–450 cm^{-1} .

For each of these intervals, a binary label is assigned to every sample. A value of "1" indicates the presence of a strong, well-defined spectral dip within the given interval, while a "0" indicates its absence. This transforms the problem into a multi-label classification task, where each sample is associated with eight binary output labels corresponding to the eight spectral regions.

To solve this classification problem, machine learning models are trained to predict all eight labels simultaneously using the same set of nine input features employed in Phase 1. These features include both the biomass category (a categorical variable indicating biomass type) and eight numerical composition variables (e.g., moisture, ash, and lignin content).

The following sections present the results of this multi-label classification approach, including key performance metrics, confusion matrices, and a comparative analysis of different models. These analyses aim to assess how well various algorithms can identify the presence or absence of spectral peaks across broad intervals, and to determine whether this simplified prediction framework provides more robust and interpretable results than full-spectrum regression.

4.2.1 Overall classification metrics

Figure 4.4 presents a comparative analysis of classification performance across four models—Logistic Regression, Random Forest, Gradient Boosting, and SVM with a RBF kernel—using Hamming Accuracy and Micro-F1 score as evaluation metrics. These metrics offer a comprehensive view of how effectively each model predicted the presence or absence of spectral peaks across the eight predefined broad intervals.

Among the models tested, Logistic Regression demonstrated the highest overall performance, achieving a Hamming Accuracy of 0.75 and a Micro-F1 score of 0.79. This indicates that the model was particularly effective at generalizing from the training data to identify peak/no-peak patterns in the test set. Notably, despite its simplicity and linear nature, Logistic Regression proved to be the most reliable model in this multi-label classification setting.

The SVM with RBF kernel followed closely, with a Hamming Accuracy of 0.68 and a Micro-F1 score of 0.75. This performance suggests that the SVM was able to model some of the underlying non-linear relationships between the input features and spectral outputs, albeit with slightly less consistency than Logistic Regression.

Random Forest produced results that were largely comparable to the SVM, with a Hamming Accuracy of 0.69 and a Micro-F1 score of 0.71. Its ability to handle feature interactions and non-linearity contributed to solid overall performance, though not sufficient to outperform the linear baseline.

Gradient Boosting, in contrast, trailed the other models with a Hamming Accuracy of 0.59 and a Micro-F1 score of 0.66. This relatively lower performance may reflect sensitivity to parameter settings or overfitting to the training data, particularly given the small dataset size.

Overall, the results depicted in Figure 4.4 suggest that Logistic Regression was best suited to this classification task, offering a favorable balance of accuracy and stability. Both SVM and Random Forest showed reasonable performance, while Gradient Boosting appeared less effective under the given constraints. These findings support the use of simpler, well-regularized models when data is limited and the task involves broad, interpretable spectral patterns.

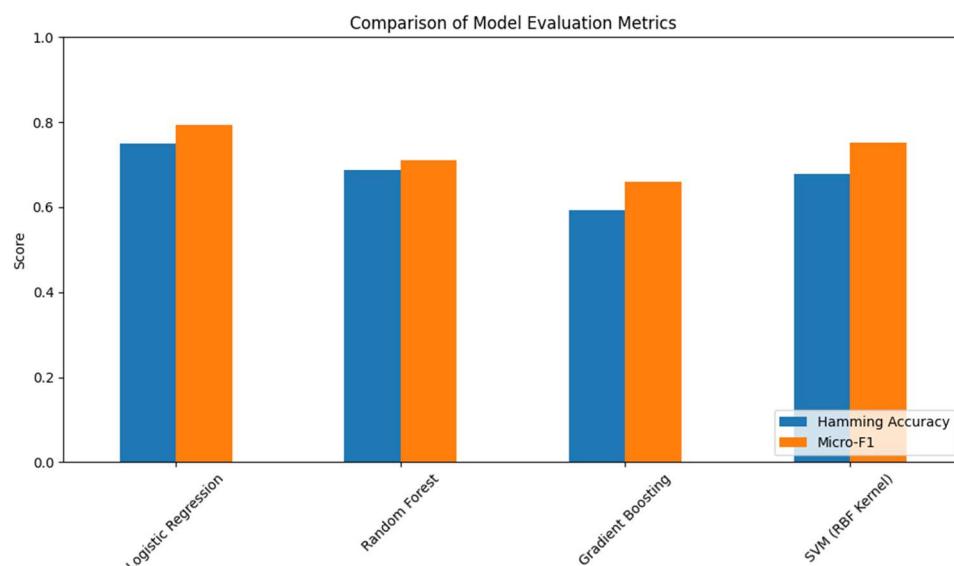


Figure 4.4 : Comparison of model evaluation metrics.

4.2.2 Confusion matrices by interval

Figure 4.5 shows the confusion matrices for the Gradient Boosting model. While the model makes some correct predictions, especially in intervals 5 and 6, its performance is inconsistent across other intervals. In interval 2, the model incorrectly predicts several "1" labels for actual "0" cases, indicating a tendency toward false positives. Similarly, interval 4 contains a mix of errors, suggesting the model may be sensitive to borderline cases or imbalanced data. These patterns highlight the model's limited generalization in complex regions.

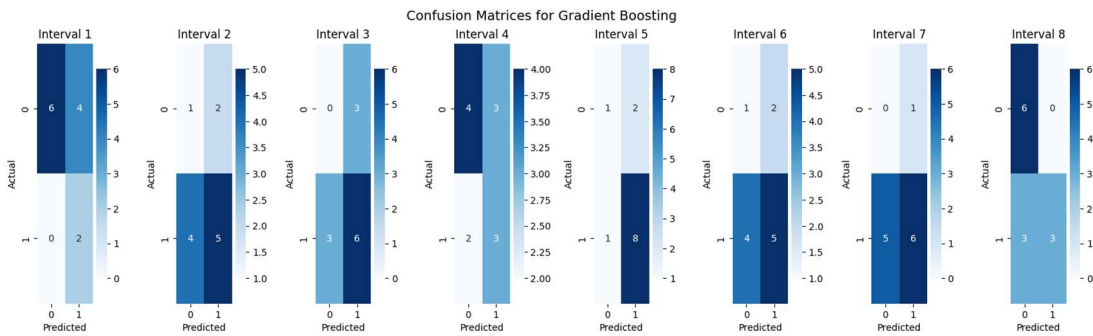


Figure 4.5 : Confusion matrices for gradient boosting.

Figure 4.6 presents the confusion matrices for Logistic Regression. The model demonstrates strong classification performance in nearly all intervals, with clearly dominant diagonal elements. In intervals 1, 3, 5, 6, and 7, Logistic Regression accurately distinguishes between peak and no-peak cases with minimal misclassifications. Interval 2 shows a few off-diagonal entries but no systematic bias, reaffirming that this model is well-calibrated and reliable across most spectral regions.

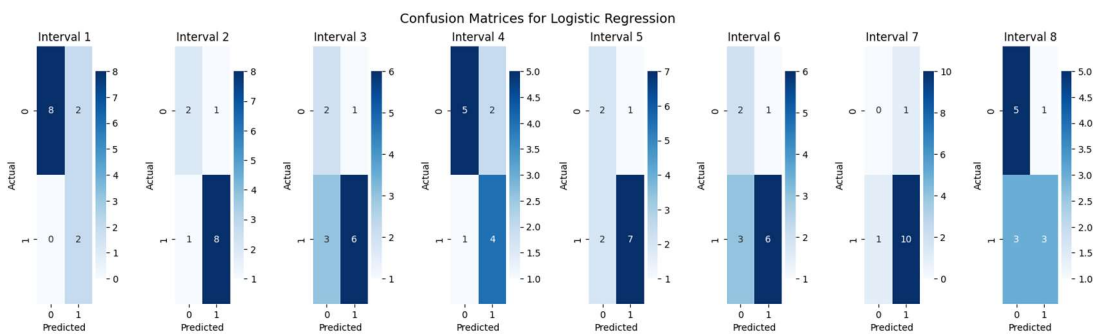


Figure 4.6 : Confusion matrices for logistic regression.

Figure 4.7 shows the confusion matrices for the Random Forest model. Performance is generally strong, with accurate predictions in intervals 3, 5, and 6. The model appears well-suited to handling non-linear relationships, though interval 2 shows a

slight bias toward false positives, and interval 4 contains a few more classification errors than other regions. Overall, the results suggest that Random Forest offers a good trade-off between flexibility and robustness.

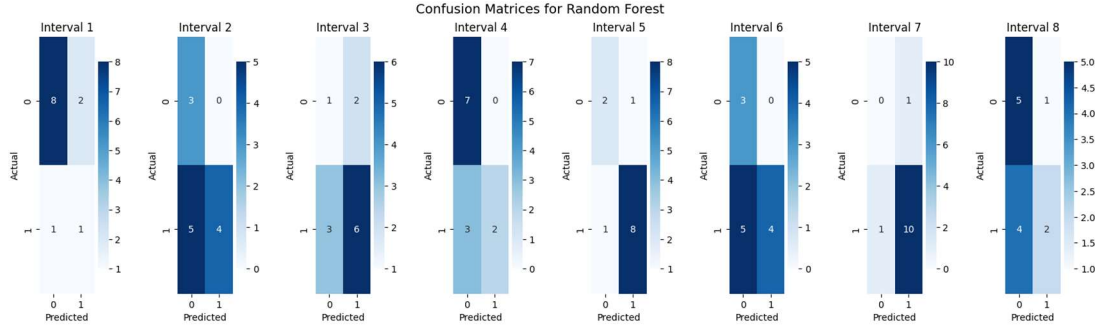


Figure 4.7 : Confusion matrices for random forest.

Figure 4.8 displays the confusion matrices for the SVM model with an RBF kernel. The model achieves consistent accuracy across all intervals, with clearly defined diagonals indicating successful classification. Intervals 1, 3, 5, 6, and 7 are particularly well-predicted. Some confusion remains in intervals 2 and 4, where a few false positives and negatives are present. Nevertheless, the overall structure of the matrices confirms that SVM performs competitively, capturing non-linear decision boundaries without significant overfitting.

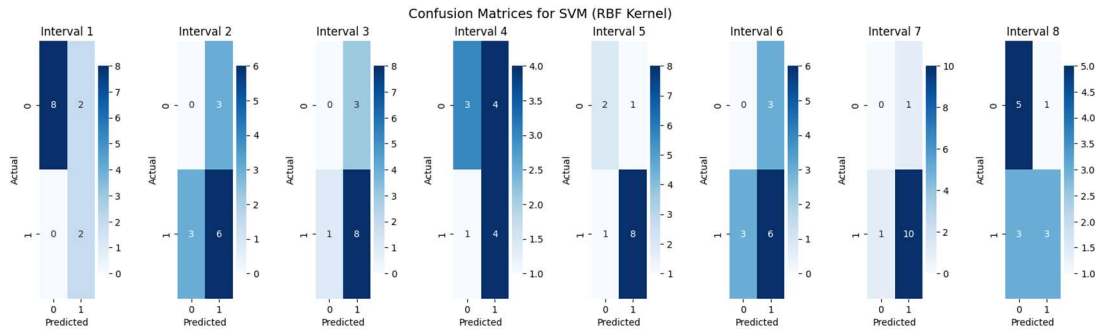


Figure 4.8 : Confusion matrices for SVM (RBF kernel).

To better understand the specific strengths and limitations of each model in Phase 2, confusion matrices were examined across the eight spectral intervals for all four classifiers: Random Forest, Logistic Regression, Gradient Boosting, and SVM. These matrices, shown in Figures 4.5 through 4.8, compare the actual labels to the predicted labels for each interval. Each label represents the presence ("1") or absence ("0") of a pronounced transmittance dip within a given spectral range. The distribution of correct

and incorrect classifications in these matrices provides valuable insight into where models perform reliably and where they struggle.

The Random Forest model offers several illustrative examples. In the first interval, spanning 4000–3700 cm^{-1} , which typically corresponds to O–H stretching vibrations, the model correctly identified the absence of a peak for most samples and made relatively few misclassifications. However, the very limited number of positive samples in this interval—those actually containing a peak—can skew the results, making the apparent performance seem stronger or weaker depending on the specific train-test split. A single false positive in such a small sample can have a disproportionate impact on the confusion matrix and derived metrics.

In the fifth interval, 1800–1500 cm^{-1} , which includes absorption features associated with carbonyl and aromatic compounds, the confusion matrix typically shows a mix of correct and incorrect predictions across both classes. If the model frequently confuses "0" with "1" in this region, it may suggest the presence of borderline cases in the dataset—samples that exhibit weak or ambiguous dips in the transmittance curve. It may also reflect an imbalance in class distribution, where one class is overrepresented and thus biases the classifier.

Additional challenges are evident in intervals seven and eight, covering the 1150–900 cm^{-1} and 900–450 cm^{-1} ranges, respectively. These regions are part of the so-called carbohydrate fingerprint zone, where peaks tend to be more subtle and numerous. In these intervals, models often default to predicting a single dominant class, especially if the training set contains few examples of the minority class. When all test samples are predicted as "0" or all as "1", it often points to class imbalance or the absence of clear signals for peak detection in the test set. In such cases, frequent misclassifications indicate that the model may not be capturing the fine-grained spectral detail necessary to distinguish peaks reliably.

The same interpretive framework applies to the confusion matrices for Logistic Regression, Gradient Boosting, and SVM. When a confusion matrix shows most predictions concentrated along the diagonal, it indicates that the model is reliably distinguishing between peak and no-peak classes. Conversely, significant numbers of false positives or false negatives—manifesting as off-diagonal elements—suggest that

the model struggles to differentiate between these categories, possibly due to overlapping features, noise, or limitations in the training data.

In sum, the confusion matrix analysis provides a more granular view of model behavior than aggregate metrics alone. It highlights the specific spectral intervals where models succeed or fail, and helps identify whether misclassifications are driven by inherent chemical ambiguity, data imbalance, or model limitations. This interval-level diagnostic is essential for evaluating not only the performance but also the interpretability and potential applicability of these models in practical spectroscopic analysis.

4.2.3 Effect of preprocessing and polynomial expansion

Several preprocessing strategies were employed to improve model performance and ensure consistent training behavior across classification tasks. One critical step was the application of standard scaling, which involved transforming all numeric input features to have zero mean and unit variance. This procedure was particularly important given the wide range of feature magnitudes—such as ash content and lignin percentage—which could otherwise disproportionately influence model behavior. All models benefited from this normalization, but the effect was especially pronounced in gradient-based methods like Logistic Regression and SVM, where proper scaling facilitated more stable convergence and improved performance.

To enhance the models' capacity to capture non-linear relationships, a second-degree polynomial feature expansion was applied to the numeric inputs. This transformation introduced interaction terms between compositional variables, enabling models to exploit potential synergies—for example, the combined influence of moisture and ash content on spectral features. The inclusion of polynomial features led to improved classification accuracy in some cases, particularly for the SVM model, which is well-suited to benefit from such higher-dimensional feature spaces. However, this expansion also introduced the risk of overfitting, especially given the relatively small dataset. Models like Gradient Boosting may have been more sensitive to these interactions, which could partially explain their comparatively lower and more variable performance across intervals.

Another key consideration in model training was the presence of class imbalance across several spectral intervals. In some regions, the number of samples labeled with a prominent peak ("1") was significantly lower than those without ("0"), leading to skewed class distributions. This imbalance can reduce classification performance and result in confusion matrices with entire rows or columns containing only zero predictions. To address this, class weighting was incorporated into the training process for Logistic Regression, SVM, and Random Forest models. By setting the "class_weight=balanced" parameter, these algorithms automatically adjusted the importance of each class based on its frequency, thereby mitigating bias toward the majority class and improving the model's ability to detect minority-class instances.

Together, these preprocessing techniques contributed to a more robust and interpretable modeling pipeline, particularly under the constraints of limited data and multi-label classification.

4.2.4 Comparison of ML models

Among the models evaluated in Phase 2, Logistic Regression achieved the strongest overall performance, attaining the highest combined Hamming Accuracy of approximately 0.75 and a Micro-F1 score of 0.79. Its success in handling the multi-label classification task is likely attributable to the simplicity and interpretability of linear decision boundaries, as well as the inclusion of balanced class weighting during training. These factors enabled it to generalize well across intervals with varying class distributions, delivering stable and consistent predictions.

Random Forest achieved a moderate level of performance, with a Hamming Accuracy around 0.69. An examination of the confusion matrices revealed that the model performed reliably in some intervals, particularly those with clearer peak signatures, but struggled in others. This variability may stem from the nature of the algorithm itself, which involves sub-sampling both data and features during tree construction. While this promotes diversity among individual trees, it may also lead to inconsistent predictions for borderline cases or underrepresented patterns in the data.

Gradient Boosting produced the lowest classification metrics among the models tested, with a Hamming Accuracy of approximately 0.59 and a Micro-F1 score of 0.66. These results suggest that, under the current parameter settings and data constraints, the model was less effective at learning the broad-interval signals. Gradient Boosting may

be more sensitive to noise or class imbalance and likely requires more extensive hyperparameter tuning or a larger dataset to capture the relevant patterns more effectively. Its relatively poor performance in several intervals also reflects potential challenges in differentiating between subtle spectral features with limited training examples.

SVM with a radial basis function (RBF) kernel achieved the second-best—or in some cases, third-best—performance, with a Hamming Accuracy of around 0.68 and a Micro-F1 score of 0.75. The use of the RBF kernel enabled the model to fit non-linear decision boundaries, making it well-suited to moderately complex classification tasks. However, its success is highly dependent on careful tuning of hyperparameters such as the regularization constant (C) and kernel coefficient (γ), which govern the flexibility and generalization capacity of the decision surface. Without such tuning, the model may either underfit or overfit certain intervals.

Overall, these comparative results highlight that simpler, regularized models like Logistic Regression can outperform more complex alternatives when working with small and imbalanced datasets. While non-linear models like Random Forest and SVM have clear advantages in modeling interactions, their effectiveness depends heavily on appropriate parameter selection and robustness to data sparsity.

4.2.5 Conclusions from Phase 2

The results from Phase 2 indicate that broad-interval classification is substantially more tractable than the full-spectrum regression task addressed in Phase 1. The observed Hamming Accuracy and Micro-F1 scores across the models are markedly higher than the coefficient of determination (R^2) values obtained in regression. This suggests that it is considerably easier for the models to make a binary determination—whether a peak is present or not within a given interval—than to predict precise intensity values at each wavenumber across the spectrum.

Among the classification models, Logistic Regression consistently emerged as the top-performing method under the conditions of this study. Its superior performance may be attributed to the nature of the classification problem itself, which is simpler and more balanced than full-spectrum regression, and to the relatively small sample size (56 observations). The linear structure and regularization properties of Logistic

Regression appear to offer an optimal balance between bias and variance in this context.

Preprocessing strategies, including normalization and polynomial feature expansion, contributed positively to model performance. Standardizing the input features ensured fair weighting across variables, while the introduction of second-order interaction terms enabled models—particularly SVM—to capture non-linear dependencies. Nevertheless, some intervals remained challenging, particularly when the class distribution was highly imbalanced or the spectral dip signal was ambiguous. In such cases, even the best-performing models exhibited reduced accuracy, reinforcing the importance of careful feature engineering and class balancing techniques.

The confusion matrices further illustrate that model performance varies across different spectral intervals. Some intervals, such as 4000–3700 cm^{-1} and 3700–3000 cm^{-1} , are heavily dominated by the "no peak" class, which simplifies classification but may obscure rare but chemically meaningful peaks. Other intervals, such as 1800–1500 cm^{-1} , tend to exhibit a more balanced distribution between classes, allowing for a more informative assessment of the models' discriminative capabilities.

Building on the insights gained from Phase 2, the next modeling stage—Phase 3—will refine the classification task by focusing on narrow, chemically specific spectral bands. These regions are more directly associated with key functional groups and structural motifs in biomass (e.g., lignin, cellulose), and thus offer the potential for higher classification accuracy and improved chemical interpretability.

4.3 Model Performance (Phase 3: Narrow-Range Classification)

In Phase 3, the classification task focuses on narrower intervals of the FTIR spectrum. Unlike Phase 2—which classified peaks in broad spectral bands—this phase aims to detect pronounced dips within three highly specific wavenumber ranges that are strongly tied to functional groups of interest (e.g., 3000–2800 cm^{-1} , 1800–1500 cm^{-1} , 1150–900 cm^{-1}). The results below show that restricting predictions to these more chemically specialized intervals often yields higher accuracy and interpretability.

4.3.1 Overall classification metrics

Figure 4.9 presents a comparison of Hamming Accuracy and Micro-F1 scores for four classification models—Logistic Regression, Random Forest, Gradient Boosting, and SVM with an RBF kernel—applied to the narrow-range classification task in Phase 3. In this phase, the focus shifts to three chemically significant spectral regions, and the results indicate a general improvement in classification performance compared to the broader-interval predictions of Phase 2.

Random Forest achieved the highest overall performance, with a Hamming Accuracy of 0.81 and a Micro-F1 score of approximately 0.89. This suggests that the model is particularly effective at detecting spectral peaks in narrower, more functionally specific regions, likely due to its ability to capture non-linear relationships and interactions within the data. Logistic Regression followed closely, with a Hamming Accuracy of 0.75 and a Micro-F1 score of 0.84. Although slightly behind Random Forest, these results remain strong and reinforce the model's robustness even when applied to refined spectral intervals.

Gradient Boosting showed moderate performance, achieving a Hamming Accuracy of 0.67 and a Micro-F1 of 0.77. While improved relative to some of its Phase 2 outcomes, this model continued to lag behind Random Forest and Logistic Regression, potentially due to its higher sensitivity to data volume and class imbalance. SVM recorded the lowest scores among the four models, with a Hamming Accuracy of 0.64 and a Micro-F1 score of 0.75. These results suggest that, although SVM can model non-linear boundaries, its effectiveness may be limited by sample size or suboptimal kernel parameterization in this narrower classification context.

Overall, the results shown in Figure 4.9 indicate that model performance generally improves when focusing on narrower and chemically meaningful spectral intervals. This suggests that limiting the classification task to well-defined regions—such as those associated with functional groups like lignin or carbohydrates—provides clearer, more learnable signals for machine learning models. Notably, Random Forest outperforms Logistic Regression in this phase, reversing the trend observed in Phase 2 and highlighting the strength of ensemble methods in capturing subtle distinctions within targeted spectral windows.

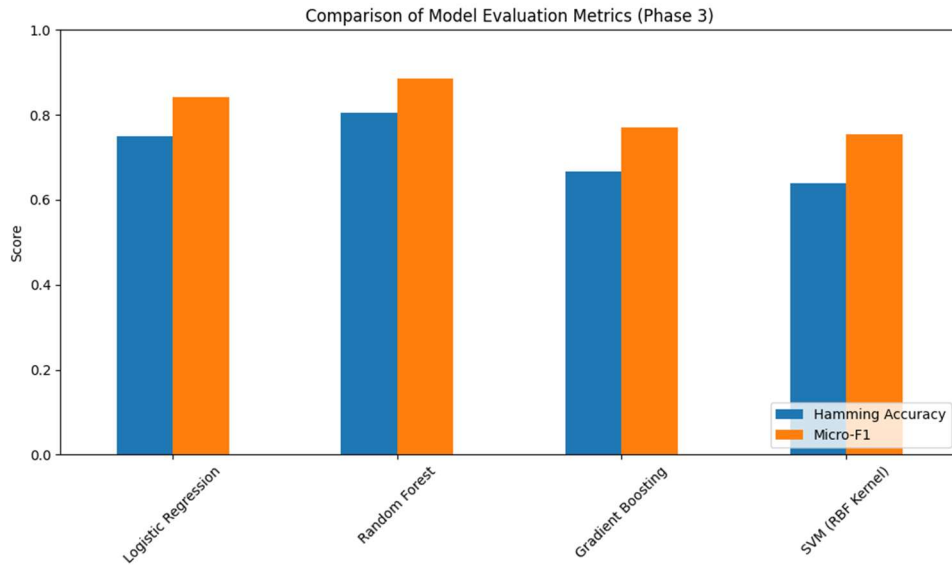


Figure 4.9 : Comparison of model evaluation metrics.

4.3.2 Confusion matrices by interval

As shown in Figure 4.10, Gradient Boosting demonstrates modest but variable performance. In Interval 1, there is a notable number of both false positives and false negatives, suggesting the model struggles to differentiate less distinct peaks. Interval 2 shows improved accuracy, with most predictions correctly classified. Interval 3 exhibits the strongest performance, where the majority of samples are accurately labeled, indicating that this spectral region provides more learnable features for the model.

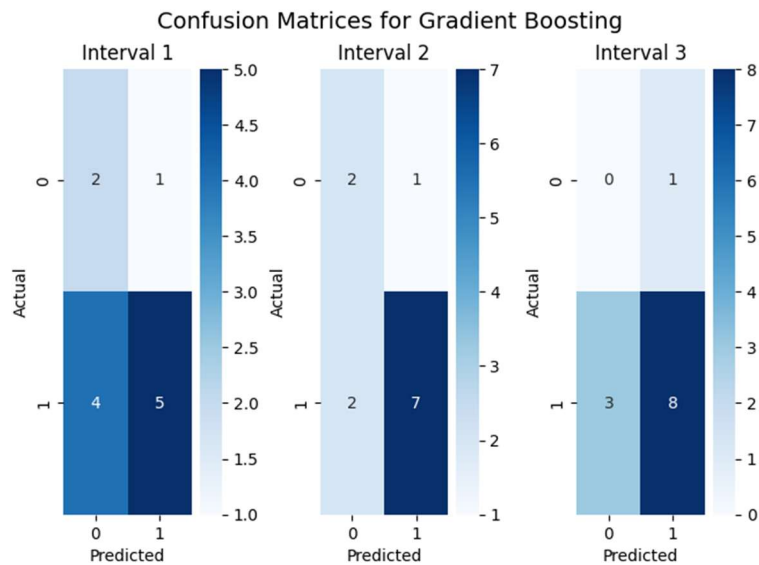


Figure 4.10 : Confusion matrices for gradient boosting.

Figure 4.11 presents the confusion matrices for Logistic Regression. The model performs well across all intervals, particularly in Intervals 2 and 3, where the diagonal dominance reflects accurate and stable classification. While Interval 1 contains a few misclassifications, the results are still balanced. These matrices support Logistic Regression’s strong performance observed in the aggregated metrics, emphasizing its effectiveness in modeling even under small data conditions.

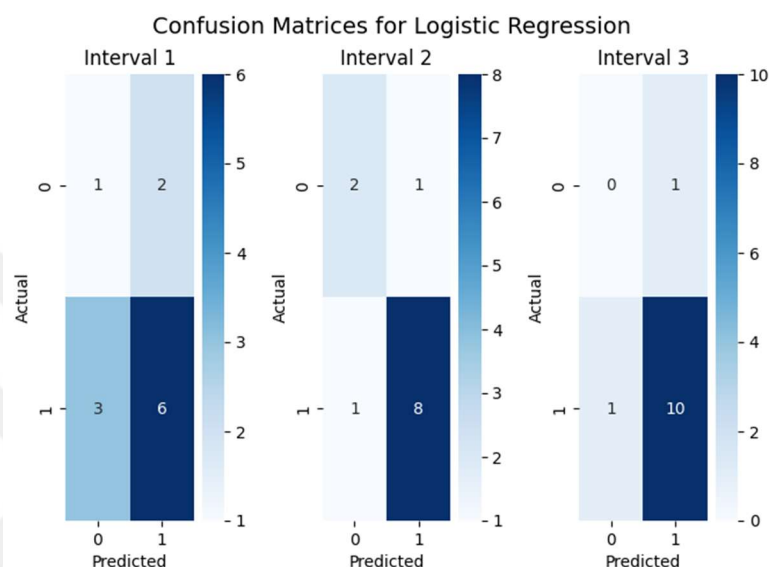


Figure 4.11 : Confusion matrices for logistic regression.

In Figure 4.12, Random Forest delivers highly consistent performance. In Intervals 2 and 3, nearly all predictions are correct, and Interval 1 shows only a small number of misclassifications. These results indicate the model's strong generalization ability in narrow, chemically relevant regions, consistent with its top Hamming Accuracy and Micro-F1 in this phase.

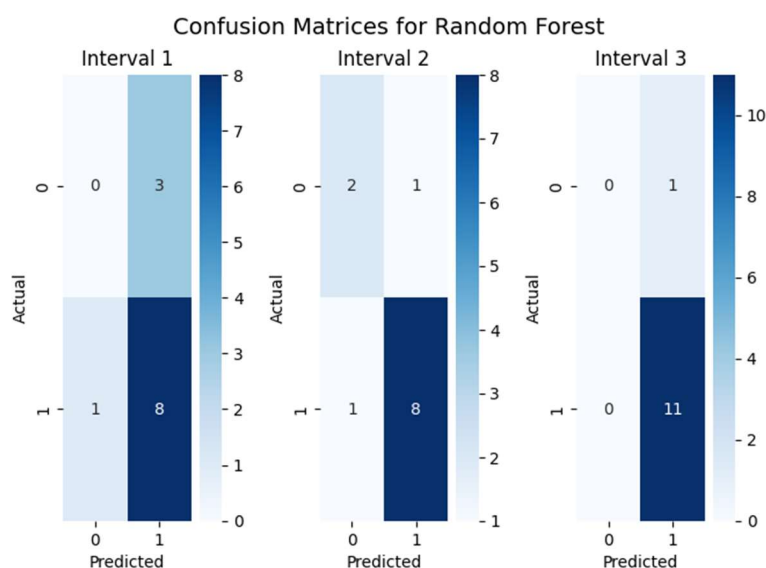


Figure 4.12 : Confusion matrices for random forest.

Figure 4.13 displays the confusion matrices for the SVM with RBF kernel. The model consistently misclassifies class "0" samples as "1". This leads to high false positive rates and reveals a significant bias toward overpredicting peak presence. While the model effectively detects actual peaks, the imbalance suggests a need for better tuning or regularization to avoid overfitting to the dominant class in the training data.

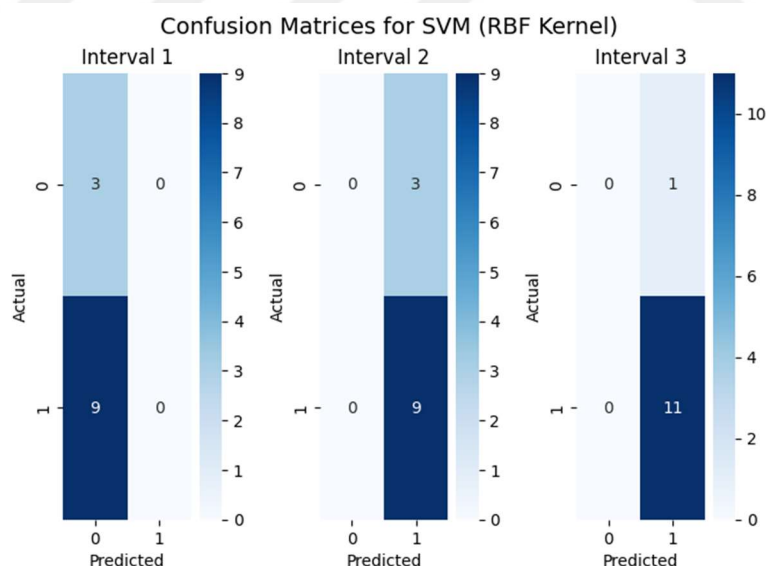


Figure 4.13 : Confusion matrices for SVM (RBF kernel).

Because Phase 3 focuses on three narrow spectral intervals, each model generates three binary outputs, indicating whether a pronounced absorbance peak is present ("1") or absent ("0") in each respective region. The corresponding confusion matrices, shown in Figure 4.10 through Figure 4.13 for Random Forest, Logistic Regression, Gradient

Boosting, and SVM respectively, provide insight into the accuracy and balance of each model's predictions. In each matrix, rows represent the actual class labels, while columns represent predicted labels, allowing a clear view of true positives, true negatives, and misclassifications.

In the first interval, corresponding to the 3000–2800 cm^{-1} range, most models demonstrate relatively low error rates. Random Forest, in particular, tends to produce very few misclassifications, with only occasional false positives or negatives. SVM, on the other hand, can struggle in this region, particularly when the class distribution is skewed. In some instances, the model may predict all samples as belonging to a single class, missing all true peaks or falsely identifying peaks where none exist.

The second interval, approximately 1800–1500 cm^{-1} , often shows better performance overall. This region typically correlates with well-defined chemical signals such as carbonyl or aromatic peaks associated with lignin. When this correlation is strong, the confusion matrices often show a high number of correct predictions for the "peak present" class. Logistic Regression frequently achieves balanced classification performance in this region, while SVM and Gradient Boosting can display more polarized behavior, such as classifying all samples into one category—particularly when training data is limited or features are less distinct.

In the third interval, covering the 1150–900 cm^{-1} range, which corresponds to the carbohydrate fingerprint region, many models exhibit improved predictive performance. This improvement is often attributed to strong alignment between certain compositional variables—such as holocellulose content—and the presence of detectable spectral dips. Random Forest, in particular, tends to show a high concentration of correct predictions along the matrix diagonal, suggesting a close relationship between input features (e.g., dry ashless holocellulose or related polysaccharide indicators) and the corresponding spectral patterns. This interval frequently yields clearer classification boundaries and better signal-to-noise characteristics, making it easier for models to learn the correct decision rules.

Careful examination of these confusion matrices helps clarify which spectral intervals are consistently predicted across models and which remain ambiguous, often due to imbalanced data or insufficient signal clarity. For example, when the positive class is rare, the SVM may entirely fail to detect it, resulting in confusion matrices with no

correct predictions for "peak = 1." This highlights the importance of both model selection and input preprocessing in narrow-band spectral classification tasks.

4.3.3 Discussion and insights

The results from Phase 3 demonstrate a clear improvement in classification performance compared to Phase 2. Models such as Random Forest and Logistic Regression achieved Hamming Accuracy values exceeding 0.75, with the best model reaching 0.81. These values surpass the majority of broad-interval scores observed in the previous phase. This improvement confirms that narrowing the focus to specific spectral intervals enhances the detectability of chemically meaningful signals. For example, well-defined peaks corresponding to lignin-related vibrations are more easily captured in a narrow window than across a generalized aromatic region, where signal dilution and overlapping features reduce discriminative clarity.

Model rankings in this phase further emphasize the benefit of tailoring algorithms to the problem's structural characteristics. Random Forest outperformed all other models, with a Hamming Accuracy of 0.81 and a Micro-F1 score near 0.89. Its ensemble nature allows it to model localized and potentially non-linear relationships that are prominent in specific wavenumber intervals. Logistic Regression also performed strongly, suggesting that when spectral intervals are well-aligned with distinct functional group signatures, even a simple linear model can yield highly accurate predictions. In contrast, Gradient Boosting and SVM delivered lower accuracy but still produced reasonable confusion matrices. These models may require more extensive hyperparameter optimization or a larger training dataset to reach the performance levels of the top classifiers.

Data preprocessing once again played a crucial role in supporting classifier performance. Standard scaling ensured that each input feature contributed proportionally to model training, while polynomial feature expansion enabled the capture of non-linear interactions between compositional variables. Because the classification task in Phase 3 targets narrow and chemically well-characterized spectral bands, the signal for "peak vs. no peak" tends to be sharper and less noisy. As a result, the models benefited from relatively simple feature transformations, provided that the compositional variables were directly linked to the expected spectral features.

Analysis of the confusion matrices provides additional insight into model behavior. In intervals that strongly correspond to identifiable chemical features—such as the carbonyl/aromatic region or carbohydrate fingerprint—the matrices typically show dominant diagonal entries, reflecting a high proportion of correct classifications. This pattern indicates that the spectral presence or absence of a peak in those bands is reliably learnable from the compositional inputs. Conversely, in cases where models such as SVM failed to make any correct predictions for one class, the underlying issue was usually an imbalanced class distribution or poor separability within the feature space defined by the model’s kernel function. Such outcomes underscore the importance of both data quality and model configuration in narrow-interval classification.

These findings carry important practical implications. The ability to accurately detect peaks in specific spectral windows, such as 1800–1500 cm^{-1} for carbonyl and aromatic groups or 1150–900 cm^{-1} for carbohydrate-related signals, is valuable for rapid chemical screening. This capability supports fast and targeted assessments of biomass composition, offering a potentially powerful diagnostic tool in bioenergy applications. The narrow-interval classification strategy explored in Phase 3 could be readily adapted as a fast-lane analytical step—used to confirm or exclude the presence of particular functional groups with high confidence, thereby providing a direct bridge between spectral data and compositional interpretation.

4.3.4 Conclusions from Phase 3

The results of Phase 3 confirm that refining the prediction task to focus on narrower spectral intervals leads to marked improvements in classification accuracy. Compared to broad-band detection in Phase 2, the models perform better when tasked with identifying the presence or absence of peaks within more chemically targeted wavenumber ranges. This improvement suggests that specific narrow-band transmittance dips exhibit a stronger and more direct correlation with underlying biomass composition features, enhancing the learnability of the classification task.

Among the models tested, Random Forest demonstrated the most consistent and accurate performance in these specialized intervals, achieving a Hamming Accuracy of approximately 0.81 and a Micro-F1 score near 0.89. Its ensemble structure allows it to effectively capture the complex, localized relationships between input features

and spectral responses, particularly in contexts where chemical specificity provides clear signals for learning.

The confusion matrices corresponding to each model further support these findings. In several intervals, particularly those associated with well-defined functional groups, the matrices reveal nearly perfect classification performance, with a high concentration of true positives and true negatives. In other intervals, however, performance varied slightly depending on the model and the distribution of samples across classes, occasionally leading to false positives or false negatives. These inconsistencies highlight the influence of both algorithmic sensitivity and dataset balance on classification reliability.

The success of narrow-range classification in Phase 3 reinforces the principle that each distinct region of the FTIR spectrum corresponds to specific chemical functionalities—such as carbohydrates, lignin, or fatty acids. By isolating these regions and modeling them independently, this phase achieved not only superior predictive accuracy but also significantly enhanced interpretability. Unlike the more generalized prediction tasks in Phases 1 and 2, the focus on targeted intervals directly tied to known chemical groups allows for clearer associations between spectral behavior and sample composition.

In conclusion, Phase 3 validates the underlying hypothesis that concentrating on chemically relevant wavenumber intervals enables more accurate and confident predictions of peak presence. This strategy represents an important step toward real-world applications, where rapid confirmation of specific chemical traits in biomass may be achieved using only a small number of carefully selected spectral windows. Such an approach offers the potential for faster, more interpretable, and practically relevant FTIR-based screening tools in industrial and bioenergy contexts.

4.4 Discussion

The results across all three modeling phases offer meaningful insights into the chemical interpretability of FTIR spectra in the context of biomass composition. Phase 1 revealed the inherent complexity of predicting full-spectrum FTIR intensities from compositional data. Despite using a variety of regression models, the resulting R^2 values remained relatively low, indicating that while certain chemical features—such

as those associated with dominant functional groups—may correlate with spectral signals, others likely involve subtle or nonlinear interactions that are not easily captured by conventional regression techniques.

In contrast, Phase 2 introduced a classification framework, reformulating the task as binary prediction of peak presence or absence within broad spectral intervals. This adjustment led to clear improvements in performance. Logistic Regression emerged as the top-performing model, likely due to its ability to learn simple, stable decision boundaries even under conditions of limited data. Nevertheless, some misclassification was observed in regions such as the 1800–1500 cm^{-1} band—commonly associated with carbonyl or aromatic vibrations—and the 1150–900 cm^{-1} carbohydrate fingerprint region. These issues are likely the result of overlapping spectral features, low signal contrast, or class imbalance.

Phase 3 further refined the approach by narrowing the classification task to highly specific, chemically relevant regions. This targeted strategy produced the highest classification accuracy of all phases, with Random Forest outperforming the other models at 81% Hamming Accuracy and a Micro-F1 score of 89%. These results strongly support the strategy of focusing on well-defined wavenumber intervals associated with distinct functional groups—such as lignin-related aromatic peaks or carbohydrate-linked regions—rather than attempting to model broad, ambiguous spectral features. The progressive improvement from Phase 1 through Phase 3 underscores the value of aligning model design with underlying chemical structure.

When placed in the context of current literature, these findings align well with broader trends in FTIR-machine learning integration. Full-spectrum regression remains a highly complex task, primarily due to the high dimensionality and inherent spectral redundancy. The relatively low R^2 values achieved in Phase 1 (~ 0.04 – 0.21) are comparable to similar efforts in the literature. For instance, a recent study applying a MLP to predict FTIR spectra from compositional data reported an R^2 near 0.21, while simpler linear models like PLS and Ridge Regression performed even lower (Kartal & Özveren, 2021).

On the other hand, models that use FTIR spectra as input to predict composition have demonstrated much stronger results. Studies by Acquah et al. (2016b), He et al. (2022b), and Xian et al. (2023) have shown that models such as PLS, Random Forest,

and Artificial Neural Networks (ANNs) can achieve R^2 values above 0.80. For example, PLS reached $R^2 = 0.956$ for cellulose prediction, while k-Nearest Neighbors models attained R^2 values as high as 0.93–0.97 for elemental analysis from ATR-FTIR spectra (Acquah et al., 2016b; He et al., 2022b; Xian et al., 2023).

Spectral classification tasks—like those employed in Phases 2 and 3—tend to be even more successful in the literature. SVM-based approaches for functional group identification have achieved over 93% accuracy using FTIR data alone (Wang et al., 2020), while multi-modal classifiers based on ANNs have reached macro-averaged F1 scores near 0.93. In applications directly related to biomass, Random Forest classifiers have been used with FTIR fingerprint regions to categorize biofuel pellets and vinegar types with accuracies exceeding 97% (Calle et al., 2021; He et al., 2022b). These results further validate the design of the phased approach used in this study—progressively moving from full-spectrum regression in Phase 1 to broad-interval classification in Phase 2, and finally to highly specific, chemically interpretable classification in Phase 3—culminating in improved model accuracy and interpretability. The data support the conclusion that peak-based or interval-specific classification approaches are more aligned with the practical and analytical needs of chemical diagnostics in biomass studies.

Despite these positive results, several limitations must be acknowledged. One of the most pressing challenges is data imbalance, particularly in Phase 3, where certain spectral intervals contain very few samples labeled with peak presence. This can bias models toward the majority class and reduce sensitivity to chemically meaningful features. Future work may address this by using synthetic oversampling techniques such as SMOTE or applying weighted loss functions during model training.

Feature engineering, while moderately explored in this study through standardization and polynomial expansion, offers further opportunities for refinement. Domain-specific transformations such as spectral derivatives, baseline correction, or dimensionality reduction through PCA may improve robustness by reducing noise and enhancing relevant patterns.

Another limitation is model interpretability. While Random Forest and Gradient Boosting yielded strong classification performance, their internal decision processes are inherently complex. Interpretable machine learning techniques—such as SHAP—

could help quantify how each compositional feature contributes to the classification of individual spectral intervals, increasing both transparency and trust in the model's predictions.

The relatively small dataset size also poses a limitation to generalizability. Although the models performed well under cross-validation, their applicability to a wider variety of biomass types or processing conditions remains uncertain. Expanding the dataset to include a broader range of species, pretreatment methods, and compositional profiles would allow for a more comprehensive validation of the models and their adaptability to real-world conditions.

Finally, while the models show promise in predicting peak presence, further validation is needed to ensure that these predictions correspond to true chemical phenomena. Cross-validation using orthogonal analytical techniques such as nuclear magnetic resonance (NMR) or mass spectrometry would provide stronger chemical evidence that the detected peaks align with the expected functional groups. This would further strengthen the practical utility of FTIR-ML methods in bioenergy and material characterization.

In summary, the study demonstrates that narrowing the spectral focus and aligning machine learning approaches with chemically meaningful intervals significantly enhances predictive performance and interpretability. These findings provide a solid foundation for further development of FTIR-based diagnostics and highlight the importance of tailoring machine learning strategies to the domain-specific characteristics of spectroscopic data.



5. CONCLUSIONS AND FUTURE WORK

5.1 Summary of Key Findings

This thesis demonstrated the viability of ML to interpret FTIR spectra for biomass characterization at three distinct levels of detail. In Phase 1, models attempted to predict the entire FTIR profile from nine compositional features, but faced the challenge of high-dimensional outputs and relatively low R^2 scores. Although MLP performed best, overall accuracy remained modest, reflecting the inherent complexity of full-spectrum regression. Moving to Phase 2, where classification targeted broad intervals of the spectrum, greatly improved performance. Logistic Regression emerged as the top performer, accurately identifying major transmittance dips (absorbance peaks) in bins such as 1800–1500 cm^{-1} or 1150–900 cm^{-1} . This phase underscored how simplifying the output to “peak present/absent” in broad wavenumber ranges can robustly capture chemical functional groups. Finally, Phase 3 focused on three narrow, functionally significant regions (e.g., 3000–2800 cm^{-1} , 1800–1500 cm^{-1} , 1150–900 cm^{-1}), achieving the highest accuracy overall. Random Forest consistently outperformed other models in these specialized intervals, confirming that zooming in on well-defined spectral windows bolsters prediction quality and interpretability.

5.2 Scientific Contributions

This study presents a multi-phase modeling strategy that progressively refines the scope of FTIR-based prediction tasks, beginning with full-spectrum regression and advancing through broad-range and narrow-range classification. This phased framework demonstrates that as the spectral focus becomes more targeted, model accuracy and interpretability improve significantly. By systematically constraining the modeling problem to increasingly specific spectral intervals, the approach reveals how the complexity of the data can be better managed and aligned with chemically meaningful structures.

The integration of modern machine learning techniques with FTIR spectroscopy forms a core contribution of this work. By applying models such as Random Forests, neural networks, and multi-label classifier chains, the study moves beyond traditional chemometric approaches like PLS and linear regression. These machine learning models successfully handle the high dimensionality and subtle variations present in biomass spectral data, validating their potential as robust alternatives for compositional estimation and spectral interpretation in biomass research.

A key insight from the analysis is the strong evidence supporting interval-specific modeling. The results clearly show that narrowing the prediction task to chemically relevant regions—particularly those associated with lignin or carbohydrate functional groups—leads to substantially better classification performance than approaches that consider the spectrum as a whole. This reinforces the idea that thoughtful selection of spectral windows is critical for building effective, interpretable FTIR-based predictive pipelines. Such strategies can focus computational and analytical efforts on the most informative regions, increasing both efficiency and accuracy.

Finally, the models and results presented in this thesis have clear practical implications for the broader field of biomass characterization. By enabling rapid, data-driven analysis of spectral data, machine learning models provide a scalable alternative to traditional wet-chemical methods, which are often time-consuming and labor-intensive. These tools can be directly applied to feedstock selection, quality control, and real-time process monitoring, offering a faster and more cost-effective route for evaluating biomass materials in industrial and research settings.

5.3 Limitations of the Study

While the results presented in this study are promising, several limitations must be acknowledged that may have influenced the outcomes and should guide future work. One key constraint was the relatively small sample size. The dataset comprised a limited number of biomass samples, which restricted the complexity and depth of models that could be employed, particularly for advanced approaches such as neural networks. With a larger and more diverse dataset—including a wider range of biomass types and processing conditions—these models could generalize more effectively and potentially outperform simpler classifiers.

A second limitation stems from class imbalance, particularly within certain spectral intervals that contained very few instances of “peak-present” cases. This imbalance occasionally led to biased classifications and reduced sensitivity to minority class patterns. Implementing data balancing techniques, such as synthetic oversampling or adjusted classification thresholds, could further enhance model performance by ensuring more equitable representation of all classes during training.

Interpretability remains another challenge, especially for high-performing but complex models such as Random Forests and neural networks. Although these models produced strong results in both broad and narrow interval classifications, their internal decision logic is not inherently transparent. Applying post hoc interpretability tools—such as SHAP or feature importance analyses—would help clarify how specific compositional variables influence model outputs, thus improving trust and transparency in predictive decision-making.

Lastly, while the narrowed spectral focus in Phase 3 improved classification accuracy, it did not fully eliminate the challenge of spectral overlap. Certain functional groups, such as hemicellulose and lignin, exhibit absorption bands that partially overlap even within tightly defined wavenumber ranges. This spectral redundancy can obscure signal clarity and complicate classification. Future research could address this issue by incorporating derivative spectroscopy, spectral deconvolution, or finer-resolution waveband analysis to better resolve subtle, overlapping peaks and enhance model sensitivity to distinct chemical signatures.

Collectively, these limitations highlight the need for ongoing refinement in both dataset design and model development to fully realize the potential of machine learning in FTIR-based biomass analysis.

5.4 Future Research Directions

Several opportunities exist to build upon the findings of this study and further enhance the performance and generalizability of machine learning models for FTIR-based biomass characterization. In terms of modeling improvements, additional feature engineering approaches may yield more refined input representations. Exploring spectral derivatives, wavelet transforms, and advanced dimensionality reduction

techniques such as auto encoders could help extract more chemically meaningful features while reducing redundancy. These methods have the potential to improve model interpretability and robustness, particularly in complex or overlapping spectral regions.

Beyond feature construction, further gains could be achieved through more extensive hyperparameter optimization and the application of advanced ensemble methods. Expanding current tuning strategies through grid search or Bayesian optimization could improve the performance of models like Random Forest and neural networks. Moreover, ensemble stacking—where predictions from multiple algorithms are combined into a unified model—may capture complementary strengths of individual classifiers and lead to higher overall classification accuracy.

Expanding the dataset to include a wider range of biomass types represents another critical area for development. Incorporating feedstocks such as agricultural residues, herbaceous grasses, and tropical hardwoods would introduce greater compositional diversity, enabling the construction of more generalized models. External validation on previously unseen samples, ideally sourced from varied geographic locations or harvested in different seasons, would be essential for evaluating model robustness in real-world settings and confirming predictive reliability beyond the training distribution.

The application of deep learning techniques to spectral data also holds significant promise. Deep neural networks, including CNNs, could be used to automatically learn and extract spectral features from raw FTIR data, potentially outperforming hand-crafted features. Additionally, sequential models such as RNNs or transformer-based architectures may be capable of modeling the wavenumber sequence itself, capturing complex dependencies across the spectral domain. This could improve classification accuracy, especially in identifying subtle or overlapping functional group signatures.

By strengthening the machine learning pipeline, broadening the diversity of the sample pool, and incorporating more sophisticated modeling architectures, future work can further improve the precision and applicability of FTIR-based classification in biomass research. The central conclusion remains clear: when applied thoughtfully, machine learning—particularly in carefully selected spectral intervals—provides a powerful

and efficient approach for extracting detailed chemical information from biomass with high accuracy and minimal experimental labor.





REFERENCES

- Acquah, G. E., Via, B. K., Fasina, O. O., & Eckhardt, L. G. (2016a). Rapid quantitative analysis of forest biomass using Fourier transform infrared spectroscopy and partial least-squares regression. *Journal of Analytical Methods in Chemistry*, 2016, 1-10. <https://doi.org/10.1155/2016/1839598>
- Bacher, A. D. (2016). *IR table*. <https://www.chem.ucla.edu/~bacher/General/30BL/IR/ir.html>
- Andrade, G. I., Barbosa-Stancioli, E. F., Mansur, A. A. P., Vasconcelos, W. L., & Mansur, H. S. (2008). Small-angle X-ray scattering and FTIR characterization of nanostructured poly(vinyl alcohol)/silicate hybrids for immunoassay applications. *Journal of Materials Science*, 43(2), 450-463. <https://doi.org/10.1007/s10853-007-1953-7>
- Apaydin Varol, E. & Mutlu, Ü. (2023). TGA-FTIR analysis of biomass samples based on the thermal decomposition behaviour of hemicellulose, cellulose and lignin. *Energies*, 16(9), 1-19. <https://doi.org/10.3390/en16093674>
- Calle, J. L. P., Ferreiro-González, M., Ruiz-Rodríguez, A., Barbero, G. F., Álvarez, J. Á., Palma, M., & Ayuso, J. (2021). A methodology based on FT-IR data combined with random-forest model to generate spectralprints for the characterisation of high-quality vinegars. *Foods*, 10(6), 1411. <https://doi.org/10.3390/foods10061411>
- Dai, F., Zhuang, Q., Huang, G., Deng, H., & Zhang, X. (2023). Infrared spectrum characteristics and quantification of OH groups in coal. *ACS Omega*, 8(19), 17064-17076. <https://doi.org/10.1021/acsomega.3c01336>
- Demirbaş, A. (2002). Relationships between heating value and lignin, moisture, ash and extractive contents of biomass fuels. *Energy Exploration & Exploitation*, 20(1), 105-111. <https://doi.org/10.1260/014459802760170420>
- Esteves, B., Sen, U., & Pereira, H. (2023). Influence of chemical composition on heating value of biomass: a review and bibliometric analysis. *Energies*, 16(10), 4226. <https://doi.org/10.3390/en16104226>
- Fadlelmoula, A., Catarino, S. O., Minas, G., & Carvalho, V. (2023). A review of machine-learning methods recently applied to FTIR spectroscopy data for the analysis of human blood cells. *Micromachines*, 14(6), 1145. <https://doi.org/10.3390/mi14061145>
- Hames, B., Ruiz, R., Scarlata, C., Sluiter, A., Sluiter, J., & Templeton, D. (2008). *Laboratory analytical procedure (LAP): preparation of samples for compositional analysis* (Issue Date 08/08/2008). National Renewable Energy Laboratory. www.nrel.gov

- He, L., Hu, W., & Wei, Y. (2022a). Lignocellulose determination and categorisation analysis for biofuel pellets based on FT-IR spectra. *Spectroscopy*, 1-13. <https://doi.org/10.56530/spectroscopy.hg8068b2>
- IR Absorption Frequencies. (2014). <https://www.eng.uc.edu/~beaucag/Courses/Characterization/IRData/IR%20Absorption%20Frequencies.pdf>
- Jabed, M. A., Kim, Y., Yarbrough, C., Harman-Ware, A. E., Olstad, J., Seiser, R., Paeper, C., Starace, A. K., & Kim, S. (2023). A machine-learning model for predicting composition of catalytic coprocessing products from molecular-beam mass spectra. *ACS Sustainable Chemistry & Engineering*, 11(32), 12055-12065. <https://doi.org/10.1021/acssuschemeng.3c01821>
- Javier-Astete, R., Jimenez-Davalos, J., & Zolla, G. (2021). Determination of hemicellulose, cellulose, holocellulose and lignin content using FTIR in *Calycophyllum spruceanum* (Benth.) K. Schum. and *Guazuma crinita* Lam. *PLOS ONE*, 16(10), e0256559. <https://doi.org/10.1371/journal.pone.0256559>
- Jesus, E., França, T., Calvani, C., Lacerda, M., Gonçalves, D., Oliveira, S. L., Marangoni, B., & Cena, C. (2024). Making wood inspection easier: FTIR spectroscopy and machine learning for Brazilian native commercial-wood-species identification. *RSC Advances*, 14(11), 7131-7143. <https://doi.org/10.1039/d4ra00174e>
- Fox, J. M. (2013). *IR handout*. https://www1.udel.edu/chem/fox/Chem333/Fall2013/Chem333Fall2013/Welcome_files/IR%20handout.pdf
- Kartal, F. & Özveren, U. (2021). An improved machine-learning approach to estimate hemicellulose, cellulose and lignin in biomass. *Carbohydrate Polymer Technologies & Applications*, 2, 100148. <https://doi.org/10.1016/j.carpta.2021.100148>
- Li, H., Chen, J., Zhang, W., Zhan, H., He, C., Yang, Z., Peng, H., & Leng, L. (2023). Machine-learning-aided thermochemical treatment of biomass: a review. *Biofuel Research Journal*, 10(1), 1170-1189. <https://doi.org/10.18331/BRJ2023.10.1.4>
- Liang, R., Chen, C., Sun, T., Tao, J., Hao, X., Gu, Y., Xu, Y., Yan, B., & Chen, G. (2023). Interpretable machine-learning-assisted spectroscopy for fast characterisation of biomass and waste. *Waste Management*, 160, 117-129. <https://doi.org/10.1016/j.wasman.2023.02.012>
- Mokari, A., Guo, S., & Bocklitz, T. (2023). Exploring the steps of infrared spectral analysis: pre-processing, (classical) data modelling and deep learning. *Molecules*, 28(19), 6886. <https://doi.org/10.3390/molecules28196886>
- NREL. (n.d.). *Biomass compositional analysis – laboratory procedures*. National Renewable Energy Laboratory. Retrieved 28 February 2025, from <https://www.nrel.gov/bioenergy/biomass-compositional-analysis.html>

- Pushpa, S. R., Awoyale, A. A., Lokhat, D., Sukumaran, R. K., & Savithri, S.** (2024). Infrared-based machine-learning models for the rapid quantification of lignocellulosic multi-feedstock composition. *Bioresource Technology Reports*, 25, 101747. <https://doi.org/10.1016/j.biteb.2023.101747>
- Segato, F., Damásio, A. R. L., de Lucas, R. C., Squina, F. M., & Prade, R. A.** (2014). Genomics review of holocellulose deconstruction by *Aspergilli*. *Microbiology & Molecular Biology Reviews*, 78(4), 588-613. <https://doi.org/10.1128/MMBR.00019-14>
- Shimadzu.** (n.d.). *Algorithms used for data processing in FTIR*. Shimadzu Corporation. Retrieved 28 February 2025, from https://www.shimadzu.com/an/service-support/technical-support/ftir/tips_and_tricks/algorithms.html
- Szymańska-Chargot, M. & Zdunek, A.** (2013). Use of FT-IR spectra and PCA to the bulk characterisation of cell-wall residues of fruits and vegetables along a fraction process. *Food Biophysics*, 8(1), 29-42. <https://doi.org/10.1007/s11483-012-9279-7>
- Tayyab, M., Noman, A., Islam, W., Waheed, S., Arafat, Y., Ali, F., Zaynab, M., Lin, S., Zhang, H., & Lin, W.** (2018). Bioethanol production from lignocellulosic biomass by environment-friendly pretreatment methods: a review. *Applied Ecology & Environmental Research*, 16(1), 225-249. https://doi.org/10.15666/aer/1601_225249
- Tkachenko, Y. & Niedzielski, P.** (2022). FTIR as a method for qualitative assessment of solid samples in geochemical research: a review. *Molecules*, 27(24), 8846. <https://doi.org/10.3390/molecules27248846>
- Wang, Z., Feng, X., Liu, J., Lu, M., & Li, M.** (2020). Functional-group prediction from infrared spectra based on computer-assist approaches. *Microchemical Journal*, 159, 105395. <https://doi.org/10.1016/j.microc.2020.105395>
- Whatley, C. R., Wijewardane, N. K., Bheemanahalli, R., Reddy, K. R., & Lu, Y.** (2023). Effects of fine grinding on mid-infrared spectroscopic analysis of plant-leaf nutrient content. *Scientific Reports*, 13, 7240. <https://doi.org/10.1038/s41598-023-33558-5>
- Xian, H., He, P., Lan, D., Qi, Y., Wang, R., Lü, F., Zhang, H., & Long, J.** (2023). Predicting the elemental compositions of solid waste using ATR-FTIR and machine learning. *Frontiers of Environmental Science & Engineering*, 17(10), 121. <https://doi.org/10.1007/s11783-023-1721-1>
- Zhuang, J., Li, M., Pu, Y., Ragauskas, A. J., & Yoo, C. G.** (2020). Observation of potential contaminants in processed biomass using Fourier transform infrared spectroscopy. *Applied Sciences*, 10(12), 4345. <https://doi.org/10.3390/app10124345>



APPENDICES

APPENDIX A: Biomass Analysis Results.



APPENDIX A

Table A.1 : Biomass analysis results.

Sample Name	Structural Analysis			Proximate Analysis			
	Extractive Substance (%)	Holocellulose (%)	Lignin (%)	Humidity (%)	Volatile Substance (%)	Ash (%)	FC (%)
Arpa küspesi (Barley meal)	11.1	64.0	18.2	8.3	76.8	6.8	8.3
Ayçekirdeği kabuğu (Sunflower hull)	16.4	53.6	29.8	8.7	76.6	2.2	12.5
Ayçiçek sapı (Sunflower stalk)	19.3	70.9	5.1	8.6	67.3	13.7	10.4
Badem kabuğu (Almond shell)	4.1	74.1	18.7	1.3	76.3	4.0	18.5
Bezelye sapı (Pea stalk)	35.2	58.6	1.5	4.0	81.2	8.4	6.4
Ceviz kabuğu (Walnut shell)	16.4	70.7	7.7	6.2	82.3	2.5	9.1
Ceviz dalı (Walnut branch)	11.4	63.8	24.0	7.7	88.2	2.8	1.3
Çam kozalağı (Pine cone)	9.9	46.5	37.4	9.3	70.8	6.3	13.8
Çay atığı (Tea waste)	37.3	43.3	6.6	5.7	67.3	6.7	20.2
Çay kafeini (Tea caffeine)	20.5	32.0	41.0	7.3	72.8	6.5	13.5
Çeltik sapı (Rice husk)	17.7	70.9	5.3	6.3	64.8	16.4	12.5
Dişbudak kabuğu (Ash wood bark)	35.9	50.9	12.2	7.1	77.0	7.6	8.3
Doğu ladini odunu (Spruce wood)	8.5	71.8	19.4	6.4	81.0	0.8	11.9
Fasulye sapı (Bean stalk)	11.5	68.9	13.9	9.0	81.0	5.8	4.3
Fındık kabuğu (Hazelnut shell)	16.2	83.0	0.1	7.5	82.6	0.6	9.3
Fındık dalı (Hazelnut branch)	14.3	64.5	20.7	9.0	73.9	0.8	16.3
Fındık zürufu (Hazelnut husk)	19.4	50.0	25.4	7.7	70.7	13.0	8.7
Fıstık çamı kozalağı (Stone pine cone)	9.6	67.0	21.1	7.6	72.9	0.7	18.8
Kakao kabuğu (Cocoa shell)	23.1	36.2	35.2	10.3	65.9	5.0	18.8
Kavak odunu (Poplar wood)	7.9	79.9	12.2	7.2	83.1	0.6	9.0
Kayısı çekirdeği (Apricot kernel)	9.6	56.9	32.6	4.0	77.9	1.0	17.1
Kayısı çekirdeği kabuğu (Apricot kernel shell)	14.9	68.8	16.0	5.9	79.1	0.5	14.5
Keçiboynuzu (Carob)	26.0	39.9	25.6	11.9	62.4	8.5	17.3
Kenevir-odunsu kısım (Hemp woody part)	19.3	69.8	8.7	7.3	76.5	1.9	14.3
Kestane kabuğu (Chestnut shell)	7.5	50.7	36.8	14.0	57.3	5.0	23.8
Kırmızı mercimek kabuğu (Red lentil shell)	7.7	63.8	27.3	10.6	68.4	1.3	19.8
Kızılcık çekirdeği (Cranberry seed)	18.7	50.4	28.4	5.5	69.5	2.5	22.5
Kiraz dalı (Cherry branch)	16.9	59.9	21.9	6.1	76.7	3.1	14.2
Kivi dalı (Kiwi branch)	9.4	68.1	20.5	3.9	79.3	2.4	14.3
Kolza (Rapeseed)	14.3	50.3	27.7	10.8	77.0	7.3	5.0
Kolza sapı (Rapeseed stalk)	8.7	53.3	33.9	4.0	71.7	9.8	14.5
Melez kavak (Hybrid poplar)	4.1	66.5	26.3	9.0	81.6	3.1	6.3
Meşe kabuğu (Oak bark)	14.2	55.6	28.4	6.3	72.1	6.4	15.2
Meşe odunu (Oak wood)	18.0	63.1	17.9	6.3	72.4	0.2	21.2
Mısır koçanı (Corn cob)	19.5	62.0	8.8	5.1	79.5	1.9	13.4
Mısır sapı (Corn stalk)	21.3	48.2	27.9	8.6	69.5	5.4	16.4

Table A.1 (continued) : Biomass analysis results.

Sample Name	Structural Analysis			Proximate Analysis			
	Extractive Substance (%)	Holocellulose (%)	Lignin (%)	Humidity (%)	Volatile Substance (%)	Ash (%)	FC (%)
Nohut sapı (Chickpea stalk)	16.5	63.5	18.1	4.9	82.9	9.2	3.0
Okaliptus kabuğu (Eucalyptus bark)	24.6	63.4	11.6	8.5	72.6	6.9	12.0
Pamuk atığı (Cotton waste)	13.1	80.5	5.2	5.9	75.7	1.8	16.5
Patlıcan sapı (Eggplant stalk)	17.5	67.4	14.3	6.9	73.8	5.6	13.7
Pirina (zeytin küspesi) (Olive pomace)	23.9	50.3	23.3	5.2	84.5	5.6	4.7
Pirinç kabuğu (Rice husk)	8.7	39.9	30.9	11.3	54.4	20.6	13.8
Sarıçam kabuğu (Pine bark)	16.0	47.9	35.0	7.3	73.2	2.5	17.0
Sarıçam odunu (Pine wood)	8.3	62.2	29.5	6.1	83.4	0.2	10.3
Sedir kabuğu (Cedar bark)	17.0	43.2	39.2	5.6	64.9	2.4	27.1
Sedir odunu (Cedar wood)	17.3	61.7	21.0	6.5	82.7	0.2	10.6
Soya küspesi (Soybean meal)	21.1	55.1	17.1	12.5	67.3	6.3	14.0
Susam kabuğu (Sesame husk)	23.1	41.6	18.1	11.3	65.3	17.2	6.3
Şeftali çekirdeği (Peach pit)	9.8	57.2	32.0	5.0	74.3	1.0	19.8
Şeftali dalı (Peach branch)	17.8	67.2	14.1	5.7	72.6	4.2	17.5
Şeftali posası (Peach pulp)	39.8	32.7	25.8	6.5	86.0	1.8	5.8
Tatlı sorgum (Sweet sorghum)	29.1	60.6	8.2	3.7	78.8	4.1	13.4
Tütün (Tobacco)	24.8	44.0	11.8	4.3	73.8	16.8	5.3
Uzun asma dalı (Long vine branch)	17.2	28.6	53.4	4.4	77.3	3.8	14.5
Üzüm çekirdeği (Grape seed)	17.8	40.0	37.5	10.0	70.0	4.8	15.3
Vişne sapı (Cherry stalk)	5.7	67.0	22.6	6.0	76.0	4.8	13.3



CURRICULUM VITAE

Name Surname : **Fahreddin Talha Sağış**

EDUCATION :

- **B.Sc.** : 2021, Izmir Institute of Technology, Faculty of Engineering, Chemical Engineering Department
- **M.Sc.** : 2025, Istanbul Technical University, Faculty of Chemical Metallurgical Engineering, Chemical Engineering Department

PROFESSIONAL EXPERIENCE:

- 01/2022 – 12/2022, Istanbul - Türkiye, Project Engineer at İhlas Holding A.Ş.
- 01/2023 – Current, Istanbul - Türkiye, Process Design Engineer at ENKA İnşaat Ve Sanayi A.Ş