An Explainable AI and LLM-Assisted Framework for Within-Household SARS-CoV-2 Secondary Transmission Risk Prediction in Sweden

MINGTONG DU
MASTER´S THESIS
DEPARTMENT OF ELECTRICAL AND INFORMATION TECHNOLOGY
FACULTY OF ENGINEERING | LTH | LUND UNIVERSITY

An Explainable AI and LLM-Assisted Framework for Within-Household SARS-CoV-2 Secondary Transmission Risk Prediction in Sweden

Mingtong Du

Department of Electrical and Information Technology
Lund University

Supervisor:
Engineering Faculty: Amir Aminifar, Lund University
Medical Faculty: Atiye Sadat Hashemi, Lund University
Jonas Björk, Lund University

Examiner: Maria Kihl, Lund University

June 4, 2026

Printed in Sweden
Tryckeriet i E-huset, Lund

Abstract

Artificial intelligence (AI) is playing an expanding role in infectious disease surveillance and intervention planning by enabling analysis of large-scale health data. One of the key challenges in this area is to characterize transmission within households and to identify factors that influence both infectivity and susceptibility. The present study focuses on the COVID-19 pandemic, which not only generated unprecedented population-level data but also emphasized household secondary transmission as a central mechanism in the spread of respiratory infectious diseases.

Predicting the household secondary transmission risk at both the household and individual levels can provide complementary evidence for targeted public health intervention. This thesis presents a multi-level explainable framework for predicting household SARS-CoV-2 secondary transmission risk, applied to the linked Swedish population and healthcare registers from the 2020 pre-vaccination period, covering 252,472 households and 608,473 individuals.

The system integrates two independent prediction tasks within a unified framework: a household-level classifier that estimates the transmission probability, and an individual-level predictor that estimates relative susceptibility scores for household members. Four classifiers (logistic regression, random forest, XGBoost, and TabPFN) are benchmarked at both levels. TabPFN, a transformer-based foundation model pre-trained on causally structured synthetic datasets via in-context learning, and XGBoost show comparable performance at both the household level (AUC-ROC: 0.645 vs. 0.641) and the individual level (0.717 vs. 0.726). The focus of this study lies not in prediction per se, but in generating interpretable insights into the determinants of household transmission. Explainability is achieved through a structured three-level KernelSHAP analysis, i.e., global population-wide attribution, subgroup stratification, and local instance-level decomposition. Finally, an LLM-based explanation agent translates model predictions and SHAP attributions into plain-language narratives. Five frontier LLMs (Claude Sonnet 4.5, GPT-5.3, Grok 4.2, DeepSeek V3.2, and Llama 4) were evaluated using a two-tier framework combining qualitative and quantitative metrics, which demonstrated the real-world application potential of LLMs.

The proposed framework integrates machine learning, explainable AI, and large language models to support more transparent and accessible public health decision-making. The code and materials for this work are publicly available at `https://github.com/mingtongdu862-dot/covid19-household-transmission`.

Popular Science Summary

When a virus enters a home, who is at risk, and can Artificial Intelligence (AI) help us answer that question in plain language?

When someone tests positive for COVID-19, the people living under the same roof face an immediate and difficult situation. They have been exposed. They are sharing a kitchen, a bathroom, the same air. Yet not every household member will get infected, and not every home that harbours an infected person will spread the virus further. Understanding which households are most likely to become outbreak clusters, and which individuals within those homes are most vulnerable, is one of the practical challenges at the heart of epidemic control.

Traditional epidemiological methods have taught us a great deal about household transmission: larger households carry more risk, older people are more susceptible, and socioeconomic disadvantage amplifies vulnerability. These findings, however, are based on relatively simple statistical models that struggle to capture the complex web of interacting factors present in large-scale national health data. During the COVID-19 pandemic, Sweden accumulated detailed records on millions of individuals: where they lived, their age and income, their medical histories, and the timing of their infections. This thesis asks whether modern artificial intelligence, applied to these population-wide data, can predict household transmission risk more accurately and more usefully than existing tools.

The answer requires solving three problems at once.

The first is **prediction**. We developed a machine learning system that works at two complementary levels: it estimates the probability that transmission will spread to at least one new person in a given household, and it identifies which specific household members are most susceptible. Using data from 252,000 Swedish households in 2020 (before vaccines were available), we tested four different AI models. The most suitable one, called TabPFN, is unusual in that it was trained not on medical data alone but on millions of artificial datasets designed to mimic the kind of cause-and-effect relationships we expect to see in the real world. This gave it a meaningful advantage, and its design opens the door for future work that could investigate causal relationships among risk factors.

The second problem is **explainability**. A prediction that cannot be explained is a prediction that cannot be trusted, especially in a public health setting. We used a technique called SHAP to open the black box of the AI model and examine which factors drove each prediction. Across both the household and individual prediction

levels, family income was the most consistently present risk factor across both prediction levels, ranking first at the individual level and second at the household level, where the age of the index case was the primary driver, reflecting the real-world constraints of overcrowded housing, limited ability to isolate a sick family member, and restricted access to protective resources. The age of the person who first introduced the virus into the household was also critical; older index cases appeared to pose a more intense and prolonged source of exposure for everyone living with them. These findings mirror what epidemiologists already know from classical studies, which tells us the AI model has learned something genuinely meaningful rather than spurious statistical noise.

The third problem is **communication**. SHAP values and probability scores are useful to an AI engineer but opaque to a public health officer or policymaker who needs to decide where to direct limited intervention resources. To bridge this gap, we connected the AI system to five state-of-the-art large language models, type of technology that powers modern AI assistants. These models were given the AI's predictions and explanations and asked to write plain-language summaries for a non-technical audience. We designed an evaluation framework to check whether the summaries were accurate, logically sound, epidemiologically correct, and appropriately framed as model outputs rather than factual diagnoses. All five models achieved perfect scores on most evaluation dimensions, with expert epidemiological review conducted on a representative subset of outputs.

The broader message of this work is that factors such as income inequality and age leave clear and consistent signatures in who is most at risk when a respiratory virus enters a home. Making them legible not just to scientists, but to the decision-makers who could act on them, is a challenge that modern AI is now well-equipped to address at a population scale.

Acknowledgments

I started working on my master's thesis a year ago, quite early compared to the typical timeline. Looking back on this journey, I want to express my sincere gratitude to everyone who supported me along the way.

First and foremost, I would like to thank my supervisor, Atiye Sadat Hashemi. Over the past year, Atiye has provided me with thoughtful guidance and countless valuable suggestions, both academically and personally. Her thoughtful mentorship has greatly shaped my research and my growth as a student.

I am also deeply grateful to my supervisor, Amir Aminifar, for his continuous support throughout this year. Amir's technical insights and advice have been invaluable, and his assistance with navigating the graduation process made everything much smoother than I could have managed alone.

My sincere thanks go to Professor Jonas Björk and the entire EPI@LUND research group. Jonas is an exceptional expert in epidemiology and data science, and his expertise opened my eyes to perspectives I never would have discovered on my own, especially coming from a different background. The members of the EPI@LUND group have been wonderfully kind and welcoming, providing not only academic collaboration but also emotional and personal support that made my research journey truly enjoyable.

Last but not least, I would like to thank my family and friends for their unwavering support throughout these two years of my master's program. Their encouragement kept me going during challenging times.

This experience will remain an unforgettable chapter in both my academic journey and my life.

Declaration

I hereby affirm that this Master thesis was composed by myself, that the work herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified; nor has it been published. Where other people's work has been used (either from a printed source, Internet, or any other source), this has been carefully acknowledged and referenced.

During the preparation of this thesis, I have used Claude AI (Sonnet 4.5 and Sonnet 4.6) to assist me in the writing process to improve language, flow, and readability. After using this tool/service, I have reviewed and edited the content as needed and I take full responsibility for the content of the whole thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Infectious disease outbreaks remain a major public health challenge, with household secondary transmission playing a central role in the spread of contact-transmitted diseases. Empirical studies consistently report that household secondary attack rates are substantially higher than those observed in non-household settings [1–3], making the household a critical site for understanding and interrupting transmission chains. The SARS-CoV-2 pandemic brought this challenge to the forefront: understanding which factors drive household secondary transmission at scale is essential for informing targeted interventions and improving future epidemic preparedness [4]. In countries with comprehensive national health registers, large-scale individual-level data covering demographic, socioeconomic, and medical histories are available to support such analysis, creating both the opportunity and the imperative for data-driven predictive tools. This chapter motivates and situates the work presented in this thesis: Section 1.1 reviews the epidemiological evidence for household transmission risk, the limitations of classical methods, and the motivation for a multi-level prediction framework. Section 1.2 traces the technical progression from machine learning models, through explainable AI, to large language models as a policy communication layer. Section 1.3 states the research objectives and contributions, and Section 1.4 outlines the remainder of the thesis.

1.1 Household Secondary Transmission and the Need for Predictive Tools

A consistent epidemiological observation across respiratory viral diseases, including influenza, respiratory virus, and SARS-CoV-2, is that the household setting constitutes a high-risk environment for transmission [2, 3]. Close and prolonged contact within confined shared living spaces dramatically amplifies exposure probability compared with transient encounters in public settings. Meta-analyses of SARS-CoV-2 household transmission have estimated secondary attack rates in the range of 17 to 19 percent for the original wild-type strain [2, 3], substantially exceeding rates observed in workplaces, schools, and other community settings [5]. These studies underline the centrality of household transmission chains in sustaining epidemic growth and motivate specific research into the mechanisms and predictors of within-household spread.

Classical epidemiological methods have generated substantial knowledge about the determinants of household transmission. Compartmental models, such as the susceptible infected recovered (SIR) framework, provide mechanistic insight into population-level dynamics [6], and secondary attack rate (SAR) estimation via logistic regression and generalised linear models has identified key risk factors including age, sex, household size, and the presence of elderly members [7, 8]. However, these approaches face notable limitations when applied to the large-scale, high-dimensional datasets that have become available through linked national health registers. Linear and additive model assumptions may fail to capture complex nonlinear feature interactions. For example, the joint effect of household composition, socioeconomic deprivation, and pre-existing comorbidities on transmission risk [9], and classical methods do not scale gracefully to massive observations with complex features.

Predicting household secondary transmission risk serves two complementary public health functions, operating at different levels of granularity. At the *household level*, estimating the probability that at least one secondary infection will occur allows authorities to stratify households by risk and direct surveillance and intervention resources accordingly, for instance through prioritised contact tracing or quarantine support. Modelling work has shown that contact tracing strategies designed to exploit household structure can meaningfully reduce epidemic growth while permitting some relaxation of broader social distancing measures [10, 11]. At the *individual level*, identifying which household members are most susceptible to infection enables finer-grained interventions, such as targeted protective measures for vulnerable members or clinically informed monitoring of high-risk individuals. Quantifying age- and health-structured heterogeneity in susceptibility within households has been shown to carry direct implications for the design of within-household protection strategies [8]. Taken together, these two levels of analysis provide a more complete evidential basis for infectious disease containment than either perspective alone. The development of a unified, multi-level predictive system capable of addressing both questions simultaneously is therefore both scientifically motivated and practically relevant.

1.2 Machine Learning, Explainability, and the Role of Large Language Models

Machine learning (ML) models offer considerable potential for overcoming the limitations of classical epidemiological approaches. Ensemble tree methods and transformer-based tabular models are capable of learning nonlinear relationships and high-order feature interactions across large-scale datasets, and have demonstrated strong predictive performance for infectious disease risk stratification [12–14]. For tabular data derived from health registers, comprising heterogeneous mixtures of demographic, socioeconomic, and clinical variable, ML classifiers provide a flexible and scalable modeling paradigm well suited to the prediction problem. Among such models, Tabular Prior-data Fitted Network (TabPFN) [15, 16], a transformer-based foundation model pre-trained on causally structured synthetic tabular datasets via in-context learning, is of particular interest: its causal pre-

training prior aligns naturally with the structure of epidemiological data, where feature–outcome relationships reflect underlying biological and social mechanisms rather than arbitrary statistical patterns.

Despite their predictive power, the well-performing ML models are generally opaque. The internal logic by which a gradient-boosted ensemble or deep tabular model arrives at a given prediction is not directly interpretable by human users [17], and this opacity constitutes a significant barrier to deployment in public health decision support, where transparency is a prerequisite for institutional trust and policy legitimacy. Integrating Explainable Artificial Intelligence (XAI) into ML pipelines has therefore become an active area of research [18]. Tools such as SHapley Additive exPlanations (SHAP) [19], which decompose model predictions into per-feature contributions grounded in cooperative game theory [20], have emerged as the dominant approach for post-hoc explanation of complex ML models in healthcare and epidemiology [21–23].

The combination of a well-performing ML predictor with SHAP-based explainability gives ML engineers a principled, quantitative account of how each feature influences model predictions. Yet this output, numerical SHAP values and feature importance rankings, presupposes a level of statistical and machine learning literacy that public health officers, epidemiologists, and policy makers may not possess. A structural bottleneck therefore remains in the translation from model output to policy-relevant evidence.

Recent advances in large language models (LLMs) present a promising mechanism for closing this communication gap. Models such as GPT-4 and open-source alternatives have demonstrated strong capabilities in technical information and generating fluent, contextually appropriate natural language [24]. These properties make LLMs well-suited to the role of an automated explanation agent: given a model prediction and its associated XAI decomposition as structured input, an LLM can generate natural language narratives that render the model's reasoning accessible to audiences without ML expertise [25, 26]. Whereas SHAP values inform an engineer about which features drove a specific prediction, an LLM-generated narrative can convey the same information in terms of actionable public health implications. This scalable, automated translation capability has the potential to substantially reduce the communication overhead between ML engineers and public health stakeholders.

1.3 Research Objectives and Contributions

This thesis develops a fully explainable, multi-level machine learning system for SARS-CoV-2 household secondary transmission risk prediction, applied to Swedish national population and healthcare registers data covering the 2020 pre-vaccination period. The system addresses two complementary prediction targets: (1) a *household-level* binary classifier estimating the probability that at least one secondary transmission event occurs within a given household, and (2) an *individual-level* susceptibility predictor assigning a susceptibility score to each household member. Both components are subjected to multi-level XAI analysis, and their outputs are passed to an LLM-based explanation agent that generates

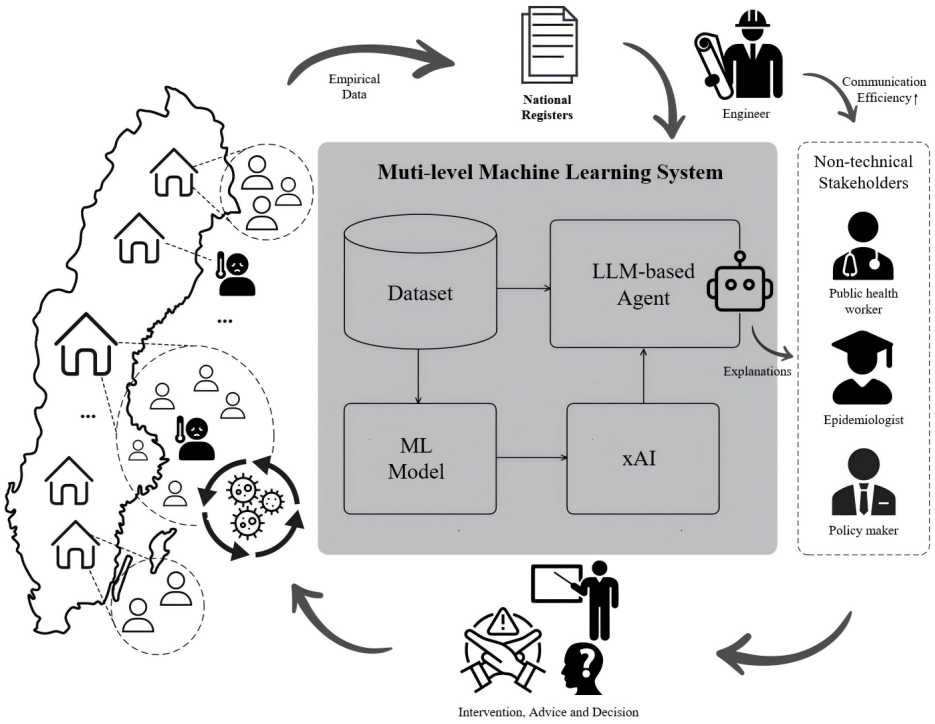


Figure 1.1: Overview of the proposed multi-level explainable ML system for SARS-CoV-2 household secondary transmission risk prediction using the Swedish national population and healthcare registers.

natural language narratives addressed to non-technical stakeholders.

Figure 1.1 illustrates this motivation and provides an overview of the complete system. Empirical disease data recorded across Swedish households are captured in national registers and fed into a multi-level machine learning system. Within that system, a trained ML model produces transmission and susceptibility predictions, which are then passed to an XAI module; the resulting feature attributions are in turn consumed by an LLM-based explanation agent that generates plain-language narratives. These narratives increase the communication efficiency between ML engineers and non-technical stakeholders, including public health workers, epidemiologists, and policymakers, who can use the insights to inform targeted interventions and decisions. The feedback arrow closing the loop from intervention back to the population underscores the practical, policy-oriented motivation of the work.

The main contributions of this thesis are as follows:

1. **Unified multi-level ML system.** We design and implement a comprehensive pipeline integrating household-level transmission probability prediction and individual-level susceptibility scoring within a single, coherent frame-

work, covering the full workflow from feature extraction and preprocessing of Swedish national register data, through model training and evaluation, to XAI analysis and LLM-based natural language explanation.

2. **TabPFN as the optimal model for register-based transmission prediction.** We benchmark multiple ML classifiers, including logistic regression, random forest, and XGBoost, and demonstrate that TabPFN [15, 16] achieves the best overall discriminative performance across both prediction levels. Specifically, TabPFN achieves the highest AUC-ROC at the household level and the highest macro F1 at the individual level. Beyond empirical performance, TabPFN is selected as the primary model because its causal generative prior aligns with the generative structure of epidemiological register data and provides a principled foundation for future counterfactual and causal inference work, while the current SHAP analysis delivers associative attributions rather than causal explanations.
3. **Multi-granularity XAI analysis.** We apply a structured three-level XAI analysis: global SHAP analysis, subgroup stratification, and local instance-level SHAP decomposition, to both the household-level and individual-level models, providing a comprehensive and epidemiologically interpretable account of the factors governing transmission risk at each level of analysis.
4. **LLM-based explanation layer with evaluation framework.** We deploy multiple LLMs as automated explanation agents for the ML system and introduce both qualitative and quantitative evaluation metrics to assess the faithfulness, comprehensibility, soundness and plausibility of the generated natural language explanations, contributing a novel evaluation framework to the emerging literature on LLMs for XAI.

1.4 Thesis Outline

The remainder of this thesis is structured as follows. Chapter 2 reviews related work spanning machine learning for infectious disease prediction, tabular foundation models, explainable AI methods, and large language models for XAI. Chapter 3 presents the methodology, covering formal problem formulation for both prediction levels, the TabPFN model and its adaptation for large-scale register data, and the XAI methods employed. Chapter 4 describes the data sources, study population, experimental setup, and results for model performance and XAI analysis at both levels. Chapter 5 presents the LLM-based explanation system, its design, and the evaluation of LLM explanation quality. Chapter 6 discusses the key findings, limitations, and directions for future work, and Chapter 7 concludes the thesis.

Chapter 2
Related Work

This chapter surveys the four bodies of literature that underpin the work presented in this thesis. Section 2.1 reviews the application of machine learning to infectious disease prediction, with particular attention to individual-level and household-level modelling of SARS-CoV-2 transmission. Section 2.2 discusses the evolution of machine learning methods for tabular data, from classical ensemble models to transformer-based architectures including TabPFN. Section 2.3 covers the development of explainable AI methods and their deployment in public health contexts. Section 2.4 reviews large language models and the emerging literature on using LLMs to generate and evaluate natural language explanations of ML system behaviour.

2.1 Machine Learning for Infectious Disease Prediction

The application of machine learning to infectious disease modeling has accelerated substantially over the past decade, driven by the increasing availability of large-scale health datasets and improvements in model performance. This section places the present work in context by reviewing the landscape of ML-based COVID-19 prediction, examining prior work at both the individual and household levels, and identifying the gap that motivates the multi-level framework developed in this thesis.

2.1.1 COVID-19 Prediction: Research Landscape

The SARS-CoV-2 pandemic generated an unprecedented volume of ML research aimed at diagnosis, prognosis, and transmission modelling [12]. Early work focused predominantly on clinical outcomes, predicting disease severity, ICU admission, or mortality from hospital records, using datasets from single or multi-centre cohorts. Ensemble tree methods, particularly Random Forest [27] and XGBoost [28], emerged as dominant approaches for structured tabular data in this domain, owing to their robustness to missing values, tolerance of heterogeneous feature types, and competitive out-of-the-box performance. As population-scale register linkage became feasible in countries with well-developed health information infrastructure, researchers began applying ML to predict infection risk at the population level,

drawing on demographic, socioeconomic, and medical history features stored in national registers [13, 14].

2.1.2 Individual-Level Infection Prediction

A substantial strand of this literature focuses on predicting whether a given individual will become infected or will experience a severe outcome, framing the problem as a binary classification task at the individual level. Willette et al. [14] used permutation-based linear discriminant analysis on UK Biobank data covering over 4,500 aged adults, identifying age, immune markers, and lipid levels as leading predictors of COVID-19 infection risk. Yang et al. [13] applied a range of ML classifiers to large-scale health records, finding that demographic and comorbidity features provided the strongest predictive signal for resistance to SARS-CoV-2 infection. Tatar et al. [23] combined XGBoost with SHAP to demonstrate that social vulnerability indicators, including housing conditions and poverty rates, were the dominant predictors of county-level infection burden in the United States, illustrating the relevance of socioeconomic context even at fine spatial and individual granularity.

These individual-level approaches share a common limitation in the present context: they typically treat each person as an independent observation, without accounting for the structured exposure environment created by household coresidence. An individual's infection risk is not determined solely by their individual characteristics but is fundamentally conditioned on the risk level of the household in which they live and on the transmission behavior of any index case present.

2.1.3 Household-Level Transmission Prediction

Household-level modeling addresses this limitation by treating the household as the unit of analysis. Classical epidemiological methods, logistic regression and generalized linear models applied to secondary attack rate (SAR) estimation, have identified age, sex, household size, and the presence of elderly or clinically vulnerable members as key determinants of within-household SARS-CoV-2 transmission [1, 7, 8]. Beyond these predominantly European and East Asian cohorts, El Baldi et al. [29] estimated a markedly higher SAR of 56.3% among household members in Morocco during the first pandemic wave, with symptomatic index cases and comorbidities in members emerging as the strongest predictors, findings that underscore the influence of healthcare-system capacity and demographic structure on intra-household transmission dynamics and motivate the inclusion of comorbidity and symptom-related features in ML-based prediction systems. Sovatzidi et al. [30] applied a fuzzy-rule-based classification framework to estimate COVID-19 transmission risk aboard cruise ships, a setting that shares many structural features with the household context: high contact rates, shared spaces, and a confined population.

2.1.4 Towards Multi-Level Frameworks

A notable gap in the existing literature is the absence of unified ML systems that address both household-level transmission probability and individual-level suscep-

tibility within a single, integrated framework. The two perspectives are epidemiologically complementary: household-level prediction identifies which households are at elevated risk of any secondary transmission event, while individual-level prediction identifies which specific household members within those households are most likely to become secondary cases. Addressing both questions simultaneously provides a richer evidential basis for targeted intervention, for example, guiding decisions about which household members should receive priority for protective measures or close monitoring, yet no prior ML study has combined these two levels of analysis within a single predictive system applied to population-wide register data.

2.2 Machine Learning for Tabular Data

Risk factor data for infectious disease prediction are typically stored in tabular form, comprising rows of individuals or households and columns of heterogeneous features spanning demographic, clinical, and socioeconomic domains. This section reviews the development of ML methods for tabular data, from classical ensemble approaches to the transformer-based models that have gained prominence in recent years, with particular attention to TabPFN as the model adopted in this thesis.

2.2.1 Classical Tabular Classifiers

Gradient-boosted decision tree (GBDT) ensembles have long represented the state of the art for tabular classification. Random Forest [27], which builds an ensemble of decorrelated decision trees via bootstrap aggregation and random feature subsampling, remains a competitive and widely used baseline. XGBoost [28], which introduces second-order gradient approximations and regularization into the boosting framework, has consistently delivered strong performance in data science competitions and applied health research. These methods offer practical advantages for epidemiological tabular data: they handle missing values without imputation, are robust to scale differences across features, and require relatively modest tuning effort compared with deep learning alternatives. Their primary limitation is that they are trained independently for each dataset and cannot leverage information from related tasks or data distributions through pretraining.

2.2.2 Transformer-Based Tabular Models and TabPFN

The application of transformer architectures, originally developed for sequential data in natural language processing, to tabular data has opened new modeling possibilities. TabNet [31], introduced by Arik and Pfister, uses a sequential attention mechanism to select which features to reason from at each decision step, providing built-in instance-wise feature selection alongside classification. Gorishniy et al. [32] proposed the Feature Tokenizer + Transformer (FT-Transformer), which embeds each tabular feature as a token and applies standard multi-head self-attention across the feature dimension; this architecture demonstrated the best average deep learning performance on a diverse set of tabular benchmarks, although gradient-boosted trees remained competitive on many individual datasets.

A more fundamental departure from the train-from-scratch paradigm was introduced by Tabular Prior-data Fitted Network (TabPFN) [15]. Rather than learning task-specific parameters, TabPFN is pre-trained offline on millions of synthetic tabular datasets generated from structural causal models (SCMs) and Bayesian neural networks, encoding a rich prior over plausible data-generating processes. At inference time, TabPFN operates via in-context learning: the training set is provided as a sequence of tokens to a transformer encoder alongside the test point, and predictions are produced without any gradient updates. This design eliminates the computational overhead of per-task training and achieves strong performance on small-to-medium datasets. Crucially, the causal pretraining prior aligns naturally with epidemiological data, where feature–outcome relationships often reflect underlying causal mechanisms, such as the effect of household size on exposure opportunity, rather than arbitrary statistical associations. TabPFN v2 [16] extended the architecture with a dedicated feature tokenisation module, support for regression alongside classification, and pretraining on approximately 130 million synthetic datasets, enabling reliable performance on datasets of up to 10,000 samples. The subsequent v2.5 release [33] further refined these capabilities through additional fine-tuning on curated real-world datasets and the introduction of specialized checkpoints for large-feature and large-sample regimes.

The interpretability of TabPFN-based predictions has received dedicated methodological attention. Rundel et al. [34] proposed and evaluated adaptations of permutation importance, partial dependence plots, and Kernel SHAP specifically designed for TabPFN's in-context learning paradigm, exploiting the model's ability to simulate retraining by modifying the context window rather than re-fitting parameters.

Despite TabPFN's strong benchmark performance, its advantages over well-tuned classical methods are not universal. A large-scale reproducible benchmark by Shaktah et al. [35], spanning twelve binary clinical prediction tasks with cohorts of up to 139,528 patients, found that TabPFN was generally competitive but exceeded the best classical ML baseline in only 16.7% of tasks, with most AUROC differences within ± 0.01. The study also reported that TabPFN incurred substantially higher computational cost, with median runtimes $5.5\times$ longer than optimized tree-based alternatives, and practical reliance on GPU acceleration. These findings suggest that the appropriate choice between TabPFN and classical ensembles is task- and dataset-dependent, and that the selection should be guided by empirical benchmarking within the target domain, as performed in this thesis.

2.3 Explainable AI and Its Applications in Public Health

As ML models have been deployed in consequential domains, including clinical decision support and public health surveillance, the demand for transparency and interpretability has grown substantially. This section traces the development of the key post-hoc explanation methods used in this thesis and reviews their application in public health and epidemiological contexts.

2.3.1 Development of XAI Methods

The field of explainable AI (XAI) encompasses both intrinsically interpretable models and post-hoc explanation techniques applied to black-box models after training. For complex models such as gradient-boosted ensembles and deep networks, post-hoc methods that operate without access to model internals are particularly valuable. Ribeiro et al. [36] introduced LIME (Local Interpretable Model-Agnostic Explanations), which approximates the black-box decision boundary in the local neighbourhood of each prediction with a simpler, interpretable surrogate model, typically a sparse linear function, fitted to perturbed samples weighted by proximity to the instance of interest. LIME established the importance of local fidelity as a desiderata for explanations, but its reliance on perturbation-based approximation introduces instability across runs, and the choice of neighborhood kernel can strongly influence the resulting attribution.

The dominant paradigm for post-hoc attribution in structured tabular data has become SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee [19]. SHAP grounds feature attribution in cooperative game theory [20]: each feature's contribution to a given prediction is computed as its average marginal contribution across all possible subsets of features, yielding attributions that satisfy desirable axioms including efficiency, symmetry, and linearity. The resulting decomposition $f(\mathbf{x}) = \phi_0 + \sum_j \phi_j(\mathbf{x})$ supports both global explanations, via the mean absolute SHAP value across the dataset, and local explanations, via per-instance waterfall or force plots. KernelSHAP [19] provides a model-agnostic approximation by formulating the Shapley value computation as a weighted linear regression over coalitions, making it applicable to any differentiable or black-box predictor at the cost of computational overhead that scales with sample size and feature dimensionality. Structuring explanations from global to local, population-wide feature importance, subgroup-level stratification, and individual-instance decomposition, has emerged as a practical framework for extracting progressively finer evidence from ML predictions [21].

2.3.2 XAI in Public Health and Epidemiology

SHAP-based explanations have been adopted across a wide range of clinical and epidemiological ML applications. Allgaier et al. [21] conducted a systematic review of XAI methods in healthcare, concluding that SHAP is the most widely adopted post-hoc explanation technique owing to its theoretical grounding and capacity to produce both global and local attributions. In the COVID-19 domain, Chadaga et al. [22] applied SHAP to explain prognosis predictions from clinical markers, identifying age and pre-existing comorbidities as the dominant drivers of severe disease outcomes. Tatar et al. [23] combined XGBoost with SHAP at the county level in the United States, demonstrating that housing conditions and poverty indicators were the most influential social determinants of infection burden.

A recurring challenge across these applications is the translation of SHAP outputs to stakeholders outside the ML engineering community. Feature importance bar charts and waterfall plots require familiarity with the concept of marginal contribution and the interpretation of positive and negative SHAP values, a level of technical literacy that public health officers, clinicians, and policy makers often do

not possess. This communication gap motivates the LLM-based explanation layer developed in this thesis.

2.4 Large Language Models

Large language models have rapidly emerged as a transformative technology with applications well beyond natural language processing. This section reviews their development, the emerging literature on using LLMs to interpret and narrate XAI outputs, and the evaluation frameworks that have been proposed to assess the quality of LLM-generated explanations.

2.4.1 Development of LLMs

The modern era of large language models began with the demonstration that scale, in model parameters, training data, and compute, yields qualitative improvements in few-shot and zero-shot task performance. Brown et al. [24] showed that GPT-3, a 175 billion parameter autoregressive transformer, could perform diverse language tasks from natural language prompts alone, without task-specific fine-tuning. Subsequent developments introduced instruction tuning and reinforcement learning from human feedback (RLHF) as alignment strategies, producing models, including GPT-4, Claude, Gemini, and open-source variants such as LLaMA, that follow complex multi-step instructions reliably and generate fluent, contextually appropriate text across diverse domains. These capabilities are directly relevant to the explanation task: an LLM that can follow structured prompts, synthesise numerical information, and generate coherent natural language is a natural candidate for translating SHAP attributions into policy-readable narratives.

2.4.2 LLMs for Interpreting XAI Outputs

The use of LLMs as automated explanation agents for ML systems is an emerging research direction. Slack et al. [26] introduced TalkToModel, an interactive dialogue system in which a language model parses user questions into executable XAI operations, such as computing SHAP values, counterfactual instances, or feature distributions, and generates natural language responses based on the results. Evaluated with healthcare workers and ML professionals, TalkToModel demonstrated substantially higher perceived usefulness and ease of use compared with conventional point-and-click explainability dashboards. This result directly supports the hypothesis that LLM-mediated explanation interfaces can bridge the communication gap between ML systems and non-technical stakeholders.

Complementary work by Zytek et al. [25] outlined a research agenda for using LLMs to generate narrative explanations from the outputs of existing XAI algorithms, treating SHAP values, feature importance scores, or partial dependence profiles as structured input to a language model rather than asking the LLM to explain its own predictions. This framing, which distinguishes LLMs as *narrators* of XAI outputs from LLMs as *self-explainers*, is the approach adopted in this thesis. The key advantages are that the factual content of the explanation is anchored to computed SHAP attributions, reducing the risk of hallucination,

while the LLM contributes the linguistic fluency and contextualization necessary to make the explanation accessible.

Agentic LLM frameworks represent a further development of this paradigm, in which multiple specialized AI agents, each backed by an LLM, collaborate to carry out complex diagnostic and explanatory workflows autonomously. Alrowais et al. [37] proposed an agentic AI framework for disease diagnosis in which agents handle multimodal clinical data processing, disease prediction via fine-tuned LLMs, and explanation generation via Retrieval-Augmented Generation (RAG) grounded in standard medical guidelines, demonstrating that LLaMA-3 achieved AUROC values of 0.88 for congestive heart failure and 0.90 for urinary tract infection prediction while producing clinically coherent explanations. This line of work illustrates the direction of travel in LLM-based medical AI: from single-model narrators of fixed XAI outputs towards orchestrated, multi-agent systems capable of adaptive, evidence-grounded explanation.

2.4.3 Evaluation of LLM-Generated Explanations

Evaluating the quality of LLM-generated natural language explanations is a non-trivial problem without established consensus. Standard NLP metrics such as BLEU and ROUGE, designed to measure surface-level n-gram overlap with reference texts, are poorly suited to the explanation setting, where multiple valid paraphrases exist and where factual faithfulness to the underlying SHAP values is more important than lexical similarity to a reference narrative. Zytek et al. [25] proposed an initial set of evaluation criteria adapted specifically for LLM-generated explanation narratives, including faithfulness (does the narrative accurately reflect the SHAP attributions?), comprehensibility (is the narrative intelligible to a non-technical audience?), and consistency (does the narrative remain stable across runs?).

Task-specific evaluation metrics have received attention in adjacent domains. Guo et al. [38] benchmarked three strategies for LLM-generated explanations of deep reinforcement learning agents, Chain-of-Thought (CoT) prompting, Monte Carlo Tree Search (MCTS) augmentation, and supervised fine-tuning, using Soundness (logical coherence of the generated reasoning) and Fidelity (alignment of the explanation with the agent's actual decision process) as primary criteria. Their finding that CoT prompting frequently produces reasoning errors while MCTS improves quality for larger models highlights the sensitivity of explanation quality to both generation strategy and model capacity, and motivates the systematic multi-LLM comparison undertaken in this thesis. Human evaluation by domain experts, in our case, public health and epidemiology specialists, provides the most direct assessment of whether explanations are both accurate and actionable for non-technical stakeholders, but is resource-intensive and does not scale easily to large numbers of generated explanations. This thesis contributes both qualitative and quantitative evaluation metrics tailored to the specific context of epidemiological transmission risk explanation, building on but extending beyond the frameworks surveyed here.

Chapter 3
Methodology

This chapter describes the methodological framework underpinning the multi-level explainable ML system developed in this thesis. Section 3.1 formalises the two prediction tasks, introduces the notation used throughout, and motivates the parallel two-model design. Section 3.2 presents all four classifiers evaluated in this thesis (logistic regression, random forest, XGBoost, and TabPFN) within a unified two-level training framework that applies each model independently to the two prediction tasks — the household-level and individual-level tasks are trained on separate datasets; the TabPFN-specific subsampling strategy and prior correction are described therein. Section 3.3 describes the explainability method SHAP, along with the three-level (global, subgroup, local) analysis structure. Finally, Section 3.4 presents the design of the LLM-based explanation system, covering the inference report format, prompt engineering strategy, and model selection.

Figure 3.1 provides a unified visual overview of the complete methodology, spanning both the training phase and the inference phase described in this chapter. In the **training phase** (top), household- and individual-level feature vectors are constructed from Swedish national registers and assembled into two separate datasets, one at the household level and one at the individual level. Four classifiers (logistic regression, random forest, XGBoost, and TabPFN) are trained independently on each dataset, and the XAI module applies global, subgroup, and local SHAP analysis to the best-performing model at both prediction levels. In the **inference phase** (bottom), the trained models produce a household-level secondary transmission probability and individual-level susceptibility scores for a selected test household; local SHAP values are computed and combined with raw feature values into a structured inference report, which is passed together with the system prompt and chain-of-thought user prompt to an LLM to generate a plain-language explanation for non-technical stakeholders.

3.1 Problem Formulation

The prediction problem addressed in this thesis involves two structurally related but analytically distinct tasks: estimating the probability that a given household will experience at least one secondary SARS-CoV-2 transmission event, and estimating the probability that a specific household member will become a secondary case (the individual susceptibility score). This section defines the case taxonomy

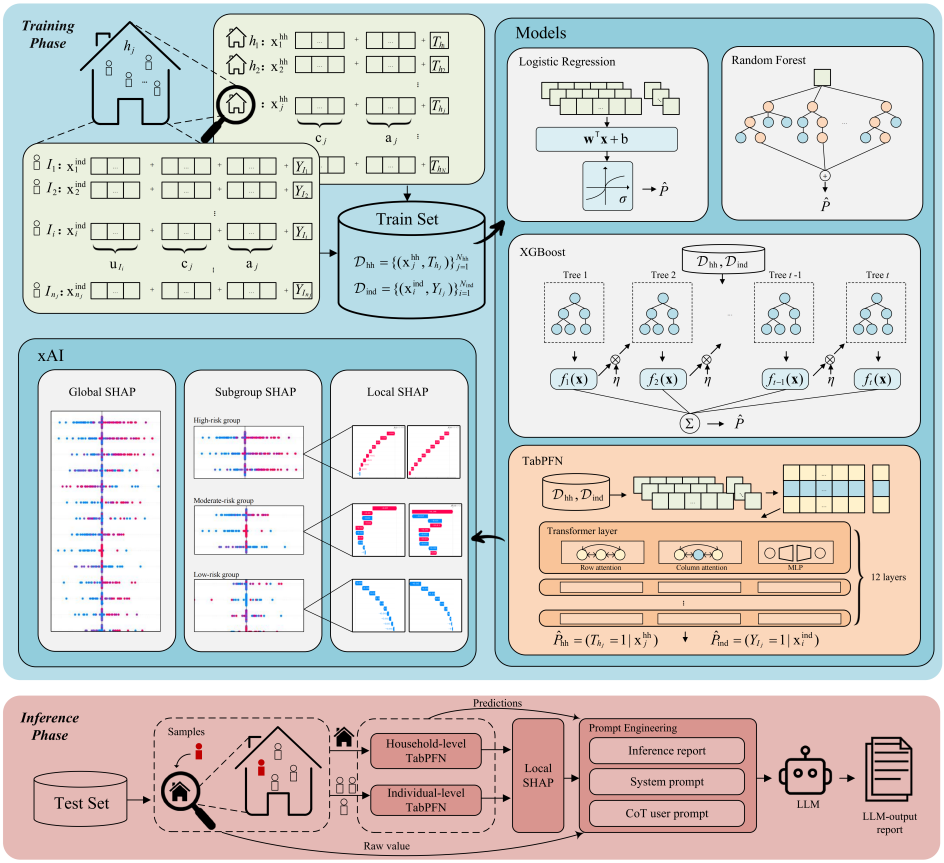


Figure 3.1: Unified overview of the methodology (training phase, top; inference phase, bottom).

and study population, introduces the unified notation used throughout, and motivates the parallel two-model design.

In this thesis, three types of input features are distinguished. $\mathbf{u}_i \in \mathbb{R}^{d_u}$ denotes *individual-specific raw features*: demographics, socioeconomic indicators, and medical history, each observed at the level of a single person I_i. $\mathbf{c}_j \in \mathbb{R}^{d_c}$ denotes *household structural features*: variables such as housing type and household size defined at the level of household h_j. $\mathbf{a}_j \in \mathbb{R}^{d_a}$ denotes *aggregated individual-level features* for household h_j, obtained by applying pooling functions $\mathcal{A}$ (e.g. mean age, gender composition) across all n_j members: $\mathbf{a}_j = \mathcal{A}(\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_{n_j}})$. These three feature types are combined into two composite input vectors: $\mathbf{x}_j^{\text{hh}}$ and $\mathbf{x}_i^{\text{ind}}$, defined precisely in Table 3.1 and Section 3.2.1. This distinction arises naturally from the register data structure; a detailed description of how these feature sets are constructed in practice is given in Section 4.1. Table 3.1 defines all principal symbols.

Table 3.1: Principal symbol definitions for Chapter 3.

Symbol	Definition
$i,\ j$	Index for an individual I_i; index for a household h_j
n_j	Number of members in household h_j
N_{hh}	Number of households in the household-level training set
N_{ind}	Number of members in the individual-level training set
$\mathbf{u}_i \in \mathbb{R}^{d_u}$	Individual-specific raw features for person I_i (demographics, socioeconomic, medical history)
$\mathbf{c}_j \in \mathbb{R}^{d_c}$	Household structural features for h_j (housing type, household size)
$\mathbf{a}_j \in \mathbb{R}^{d_a}$	Aggregated individual-level features for h_j: $\mathbf{a}_j = \mathcal{A}(\mathbf{u}_{i_1}, \ldots, \mathbf{u}_{i_{n_j}})$
$\mathbf{x}_j^{\text{hh}} \in \mathbb{R}^{d_{\text{hh}}}$	Household input vector: $[\mathbf{a}_j;\ \mathbf{c}_j]$; input to f_{hh}
$\mathbf{x}_i^{\text{ind}} \in \mathbb{R}^{d_{\text{ind}}}$	Individual input vector: $[\mathbf{u}_i;\ \mathbf{a}_{j(i)};\ \mathbf{c}_{j(i)}]$, where $j(i)$ is the household of individual i; input to f_{ind}
$Y_{I_i} \in \{0, 1\}$	Individual outcome: 1 if secondary case, 0 otherwise
$T_{h_j} \in \{0, 1\}$	Household outcome: 1 if $\geq$1 secondary transmission, 0 otherwise
$f_{\text{hh}},\ f_{\text{ind}}$	Household-level and individual-level classifiers
$\hat{P}_{\text{hh}}$	Estimated probability of $\geq$1 secondary transmission in h_j
$\hat{P}_{\text{ind}}$	Estimated probability that individual I_i becomes a secondary case
$\hat{p},\ \hat{p}_*$	Scalar predicted probability; $\hat{p}_*$ for a specific test point
$\mathcal{D}_{\text{tr}}$	Training dataset for a given cross-validation fold
$r \in (0, 1)$	Positive-class oversampling ratio in TabPFN training subsampling
$\hat{p}_{\text{corr}}$	Prior-corrected probability for TabPFN (Eq. 3.15)
$\mathcal{F} = \{1, \ldots, d\}$	Full feature index set; $d = d_{\text{hh}}$ or d_{ind} depending on prediction level
$S \subseteq \mathcal{F}$	Feature coalition; $\bar{S} = \mathcal{F} \setminus S$ its complement
ϕ_0	SHAP baseline: expected model output $\mathbb{E}_{\mathbf{x}}[f(\mathbf{x})]$
$\phi_j(\mathbf{x})$	SHAP value of feature j: signed marginal contribution to the prediction deviation from ϕ_0 (Eq. 3.17)
$v_{\mathbf{x}}(S)$	SHAP value function: $\mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})]$, the expected output with S observed and $\bar{S}$ marginalised

3.1.1 Secondary Transmission Definition and Study Population Construction

Secondary transmission definition. To define labels accurately and construct dataset at both levels, we first examine the infection timeline within each household at the individual level. As illustrated in Figure 3.2a, within any house-

hold containing infected members, individuals are infected at different points in time. Each individual's *index date* is the date of their first confirmed COVID-19 diagnosis; this date anchors the construction of that individual's medical history. For household members who were never infected, we assign their index date to be the last index date observed among the infected members of their household. This ensures that healthy members' medical histories are traced over the same outbreak window as the household's last recorded infection, capturing their health profile under active exposure conditions.

A household's infection timeline may contain more than one outbreak wave. Following WHO guidance [39], if the interval between two consecutive cases exceeds 14 days, those cases are considered to belong to separate waves. Within a single wave, the earliest case(s) are designated *index cases*. Subsequent cases are designated *secondary cases* if their infection date falls more than 2 days after the index cases' index date. This 2-day serial interval threshold [40] distinguishes coincident infections (counted as index cases) from genuine secondary transmissions. Both models draw on this individual-level case taxonomy: the household model aggregates it into the household-level outcome T_{h_j}, while the individual model uses it directly as the per-person outcome Y_{I_i}.

Our analysis focuses exclusively on each individual's first infection and each household's first outbreak wave. Individuals infected in later waves are not considered in either model.

Study population eligibility and label construction. Not all households in the registers are informative for the prediction tasks. As illustrated in Figure 3.2b, the eligible study population consists of households in which secondary transmission is structurally possible: households containing at least one index case *and* at least one member who was not co-infected during the first wave. Two categories are excluded from both datasets. First, households in which no member was infected at all provide no transmission signal. Second, households in which *all* infected members are index cases are excluded: since no susceptible member remains, secondary transmission is impossible by definition and these households yield no informative outcome contrast for either model.

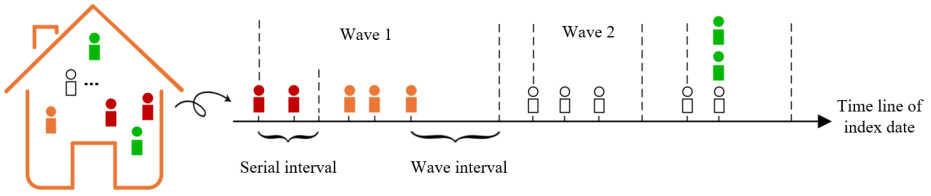
Within the eligible population, the binary outcomes are defined as follows. The household label $T_{h_j} = 1$ if at least one household member becomes a secondary case; $T_{h_j} = 0$ if no secondary transmission occurs. At the individual level, an individual I_i receives $Y_{I_i} = 1$ if they become a confirmed secondary case, and $Y_{I_i} = 0$ otherwise.

3.1.2 Household-Level Prediction Task

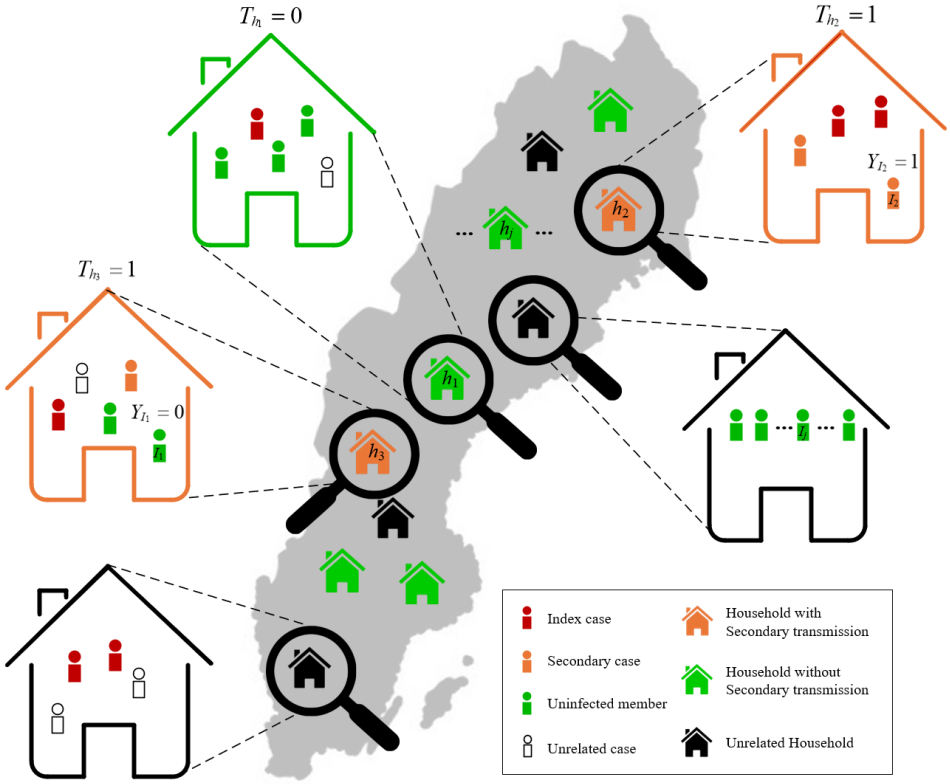
The household-level task is formulated as a binary classification problem. Given a household h_j characterised by its composite input vector $\mathbf{x}_j^{\text{hh}} = [\mathbf{a}_j; \mathbf{c}_j] \in \mathbb{R}^{d_{\text{hh}}}$, the objective is to estimate:

$$\hat{P}_{\text{hh}}(T_{h_j} = 1 \mid \mathbf{x}_j^{\text{hh}}), \tag{3.1}$$

the probability that at least one secondary transmission event will occur in household h_j. The binary label $T_{h_j} = 1$ if any member other than the index case(s) is



(a) Infection timeline within a single household.



(b) Study population eligibility and label construction.

Figure 3.2: Case definitions and label construction. In (a), infection timelines and wave demarcation within a single household. In (b), study population eligibility and binary label assignment.

subsequently confirmed as a secondary case within the defined observation window; $T_{h_j} = 0$ otherwise. The training dataset at this level is denoted $\mathcal{D}_{\text{hh}} = \{(\mathbf{x}_j^{\text{hh}}, T_{h_j})\}_{j=1}^{N_{\text{hh}}}$, where N_{hh} denotes the number of households. The model's objective is to learn $f_{\text{hh}} \colon \mathbb{R}^{d_{\text{hh}}} \to [0,1]$ such that $f_{\text{hh}}(\mathbf{x}_j^{\text{hh}})$ provides an estimate of $P(T_{h_j} = 1 \mid \mathbf{x}_j^{\text{hh}})$.

3.1.3 Individual-Level Prediction Task

While the household model addresses the question of whether transmission will occur at all, the individual-level task addresses who within the household is most at risk. Specifically, given an individual I_i characterised by their composite input vector $\mathbf{x}_i^{\text{ind}} = [\mathbf{u}_i; \mathbf{a}_{j(i)}; \mathbf{c}_{j(i)}] \in \mathbb{R}^{d_{\text{ind}}}$, where $j(i)$ denotes the household to which individual i belongs, the objective is to estimate:

$$\hat{P}_{\text{ind}}(Y_{I_i} = 1 \mid \mathbf{x}_i^{\text{ind}}), \tag{3.2}$$

the probability that individual I_i becomes a secondary case. The binary label $Y_{I_i} = 1$ if the member is subsequently confirmed as a secondary case within the defined observation window; $Y_{I_i} = 0$ otherwise. The training dataset at this level is denoted $\mathcal{D}_{\text{ind}} = \{(\mathbf{x}_i^{\text{ind}}, Y_{I_i})\}_{i=1}^{N_{\text{ind}}}$, where N_{ind} denotes the number of household members. The model's objective is to learn $f_{\text{ind}} \colon \mathbb{R}^{d_{\text{ind}}} \to [0, 1]$ such that $f_{\text{ind}}(\mathbf{x}_i^{\text{ind}})$ provides an estimate of $P(Y_{I_i} = 1 \mid \mathbf{x}_i^{\text{ind}})$.

The two prediction tasks are formulated and trained independently: f_{hh} and f_{ind} are fitted on separate datasets $\mathcal{D}_{\text{hh}}$ and $\mathcal{D}_{\text{ind}}$, share no parameters, and neither model's output is used as an input to the other. The household-level prediction $\hat{P}_{\text{hh}}$ and the individual-level susceptibility score $\hat{P}_{\text{ind}}$ are produced in parallel at inference time and presented jointly to the LLM explanation layer.

3.2 Prediction Models

This section presents the four classifiers evaluated in this thesis within a unified two-level training framework. Each classifier is trained and evaluated independently twice: once for the household-level prediction task (estimating $\hat{P}_{\text{hh}}$) and once for the individual-level prediction task (estimating $\hat{P}_{\text{ind}}$), yielding eight models in total (four classifiers $\times$ two prediction levels). The models within each level share the same cross-validation splits, training pool, and evaluation protocol, enabling direct cross-model comparison at each level.

3.2.1 Two-Level Input Representation

The two prediction tasks differ not only in their outcome variable but in the input feature vector supplied to each classifier. Drawing on the three feature types defined in Section 3.1 and Table 3.1, we define two composite input vectors:

- **Household-level input.** $\mathbf{x}_j^{\text{hh}} = [\mathbf{a}_j; \mathbf{c}_j] \in \mathbb{R}^{d_{\text{hh}}}$ concatenates the aggregated individual-level features $\mathbf{a}_j$ and the household structural features $\mathbf{c}_j$. It characterises household h_j as a whole and serves as the input to any classifier f_{hh} targeting the household outcome T_{h_j}.
- **Individual-level input.** $\mathbf{x}_i^{\text{ind}} = [\mathbf{u}_i; \mathbf{a}_{j(i)}; \mathbf{c}_{j(i)}] \in \mathbb{R}^{d_{\text{ind}}}$ concatenates the individual's own raw features $\mathbf{u}_i$, the household aggregate $\mathbf{a}_{j(i)}$, and the household structural features $\mathbf{c}_{j(i)}$, where $j(i)$ denotes the household to which individual i belongs. It characterises individual I_i in household context and serves as the input to any classifier f_{ind} targeting the individual outcome Y_{I_i}.

Throughout the model descriptions below, $\mathbf{x}$ denotes a generic input vector representing $\mathbf{x}_j^{\text{hh}}$ when describing the household-level model and $\mathbf{x}_i^{\text{ind}}$ when describing the individual-level model. Similarly, $Y \in \{0, 1\}$ represents the relevant binary outcome (T_{h_j} or Y_{I_i}), d the corresponding feature dimensionality (d_{hh} or d_{ind}), and N the level-specific training set size (N_{hh} for the household model, N_{ind} for the individual model, as defined in Table 3.1). N_+ and N_- denote the number of positive and negative training examples at the corresponding level, respectively.

3.2.2 Logistic Regression

Logistic regression models the posterior probability of the positive class as a linear function of the input features passed through a sigmoid transformation. Given weight vector $\mathbf{w} \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$, the predicted positive-class probability is:

$$\hat{P}(Y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + \exp\big(-(\mathbf{w}^\top \mathbf{x} + b)\big)}, \tag{3.3}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function. At the household level, $\hat{P}(Y = 1 \mid \mathbf{x}) = \hat{P}_{\text{hh}}(T_{h_j} = 1 \mid \mathbf{x}_j^{\text{hh}})$; at the individual level it equals $\hat{P}_{\text{ind}}(Y_{I_i} = 1 \mid \mathbf{x}_i^{\text{ind}})$. Equivalently, the model is linear in the log-odds space:

$$\log \frac{\hat{P}(Y = 1 \mid \mathbf{x})}{\hat{P}(Y = 0 \mid \mathbf{x})} = \mathbf{w}^\top \mathbf{x} + b. \tag{3.4}$$

Parameters are estimated by minimising the ℓ_2-regularised negative log-likelihood (cross-entropy loss):

$$\mathcal{L}(\mathbf{w}, b) = -\sum_{i=1}^{N} \Big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \Big] + \lambda_{\text{LR}} \|\mathbf{w}\|_2^2, \tag{3.5}$$

where $\hat{p}_i = \sigma(\mathbf{w}^\top \mathbf{x}_i + b)$ is the predicted probability for instance i, and $\lambda_{\text{LR}} > 0$ is the regularisation strength. Class imbalance is addressed by assigning inverse-frequency sample weights: each positive example receives weight $w_+ = N/(2N_+)$ and each negative example weight $w_- = N/(2N_-)$. Two independent logistic regression models are fitted: one on $\{\mathbf{x}_j^{\text{hh}}, T_{h_j}\}$ and one on $\{\mathbf{x}_i^{\text{ind}}, Y_{I_i}\}$.

3.2.3 Random Forest

Random Forest [27] is a bagging ensemble of B unpruned binary decision trees. Each tree τ_b is trained on a bootstrap sample of the training data, and at each internal node only a random subset of $q = \lfloor \sqrt{d} \rfloor$ features is considered as candidate splits, de-correlating the individual trees. The predicted positive-class probability for a new instance $\mathbf{x}$ is the average of the per-tree leaf-fraction estimates:

$$\hat{P}_{\text{RF}}(Y = 1 \mid \mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \tau_b(\mathbf{x}), \tag{3.6}$$

where $\tau_b(\mathbf{x})$ is the proportion of positive training examples in the leaf of tree τ_b that contains $\mathbf{x}$.

Each split is chosen to maximise the reduction in Gini impurity. For an internal node ν with class-frequency vector (p_0, p_1), the Gini impurity is:

$$\text{Gini}(\nu) = 1 - \sum_{c \in \{0,1\}} p_c^2. \tag{3.7}$$

A candidate split partitioning ν into children ν_L and ν_R is scored by the weighted impurity decrease $\Delta G = \text{Gini}(\nu) - \frac{|\nu_L|}{|\nu|}\text{Gini}(\nu_L) - \frac{|\nu_R|}{|\nu|}\text{Gini}(\nu_R)$, and the feature–threshold pair with the highest ΔG is selected. As with logistic regression, inverse-frequency class weights (w_+, w_-) are applied to counteract class imbalance. The ensemble size is fixed at $B = 200$ trees. Two independent Random Forest ensembles are fitted: one on $\{\mathbf{x}_j^{\text{hh}}, T_{h_j}\}$ and one on $\{\mathbf{x}_i^{\text{ind}}, Y_{I_i}\}$.

3.2.4 XGBoost

XGBoost [28] is a gradient-boosted decision tree (GBDT) ensemble that builds an additive model iteratively. At boosting round t, a new regression tree f_t is added to the ensemble, so the cumulative prediction after t rounds is:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta\, f_t(\mathbf{x}_i), \tag{3.8}$$

where $\hat{y}_i^{(t)}$ is the predicted log-odds for instance i at round t, and $\eta \in (0, 1]$ is the learning rate (shrinkage factor). The tree f_t is fitted to minimise a regularised objective obtained via second-order Taylor expansion of the loss around the predictions of round $t-1$:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{N} \Big[g_i\, f_t(\mathbf{x}_i) + \tfrac{1}{2}\, \delta_i\, f_t(\mathbf{x}_i)^2 \Big] + \Omega(f_t), \tag{3.9}$$

where $g_i = \partial_{\hat{y}} \ell(y_i, \hat{y}_i^{(t-1)})$ and $\delta_i = \partial_{\hat{y}}^2 \ell(y_i, \hat{y}_i^{(t-1)})$ are the first- and second-order gradients of the per-sample loss function ℓ (binary cross-entropy for classification). The tree complexity penalty is:

$$\Omega(f_t) = \gamma\, K_t + \tfrac{1}{2}\, \lambda_{\text{XGB}} \sum_{k=1}^{K_t} s_k^2, \tag{3.10}$$

where K_t is the number of leaves in tree f_t, s_k is the output score of leaf k, $\gamma \geq 0$ is a minimum-loss-reduction threshold for node splitting, and $\lambda_{\text{XGB}} > 0$ is an ℓ_2 regularisation coefficient on leaf scores (distinct from the logistic regression regularisation λ_{LR} in Eq. (3.5)). Class imbalance is handled by the `scale_pos_weight` parameter, set to N_-/N_+, which rescales the gradient of positive examples proportionally. Two independent XGBoost models are fitted: one on $\{\mathbf{x}_j^{\text{hh}}, T_{h_j}\}$ and one on $\{\mathbf{x}_i^{\text{ind}}, Y_{I_i}\}$.

3.2.5 TabPFN

TabPFN (Tabular Prior-Fitted Network) [15, 16, 33] is a Transformer-based in-context learning model pre-trained on a large distribution of synthetic tabular

datasets, which requires no gradient updates at inference time and achieves competitive performance on small-to-medium tabular classification tasks without task-specific hyperparameter tuning. Both the household-level and individual-level TabPFN models share the same pre-trained architecture but are applied independently to their respective input vectors ($\mathbf{x}_j^{\text{hh}}$ and $\mathbf{x}_i^{\text{ind}}$) using separate training subsamples. This subsection describes the model architecture, the balanced subsampling strategy required to scale TabPFN to national register data, and the prior correction applied when evaluating at the natural class distribution.

Model Architecture and Pre-Training

TabPFN (Tabular Prior-Fitted Network) [15, 16] is a Transformer encoder pre-trained on millions of synthetic tabular datasets generated from structural causal models (SCMs) and Bayesian neural networks. Rather than learning a task-specific mapping, TabPFN learns a *generic learning algorithm*: during pre-training it is exposed to an extremely diverse distribution of data-generating processes, encoding a rich prior over plausible feature–outcome relationships. At inference time, the model executes this learned algorithm via *in-context learning*: no gradient updates are performed.

This generative prior is particularly well-suited to epidemiological register data, where feature–outcome relationships are expected to reflect genuine causal structure (e.g. the effect of household crowding on exposure opportunity) rather than arbitrary statistical associations. The causal generative prior of TabPFN's pretraining data aligns with this expectation in a way that purely correlational models cannot guarantee.

At inference time, given a training set $\mathcal{D}_{\text{tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_{\text{tr}}}$ and a test point $\mathbf{x}_*$, TabPFN constructs an input sequence:

$$\mathcal{S} = \big[(\mathbf{x}_1, y_1),\, \ldots,\, (\mathbf{x}_{N_{\text{tr}}}, y_{N_{\text{tr}}}),\, (\mathbf{x}_*, ?)\big]. \tag{3.11}$$

Each sample is represented as a single token: feature vector $\mathbf{x}_i$ and label y_i are linearly projected into $\mathbb{R}^{d_m}$ and summed to form $\mathbf{z}_k \in \mathbb{R}^{d_m}$; for the test token the label embedding is replaced by a learnable mask $\mathbf{e}_{\text{mask}}$. A role encoding $\mathbf{r}_k \in \mathbb{R}^{d_m}$ distinguishes training tokens from the test token: $\mathbf{z}'_k = \mathbf{z}_k + \mathbf{r}_k$.

The sequence $[\mathbf{z}'_1, \ldots, \mathbf{z}'_{N_{\text{tr}}}, \mathbf{z}'_*]$ is processed by L Transformer encoder layers. Each layer applies multi-head self-attention followed by a position-wise feed-forward network. The attention output for a single head at layer ℓ is:

$$\text{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\Big)\mathbf{V}, \tag{3.12}$$

where $d_k = d_m/H$ is the per-head key dimension and H is the total number of attention heads. Multi-head attention runs H such operations in parallel over distinct linear projections and concatenates the outputs, enabling the model to capture associations between training examples and the test point from multiple representational perspectives simultaneously.

After L layers, the row of the output matrix corresponding to the test token, $\mathbf{h}_*^{(L)} \in \mathbb{R}^{d_m}$, encodes discriminative information distilled from the entire training

context. A linear classification head produces the predicted class probabilities:

$$\hat{\mathbf{p}}_* = \text{softmax}\big(\mathbf{W}_c\, \mathbf{h}_*^{(L)} + \mathbf{b}_c\big) \in \mathbb{R}^2, \tag{3.13}$$

where $\mathbf{W}_c \in \mathbb{R}^{2\times d_m}$ and $\mathbf{b}_c \in \mathbb{R}^2$ are learnable parameters fixed after pre-training. The predicted probability of the positive class is $\hat{p}_* = [\hat{\mathbf{p}}_*]_1$, corresponding to $\hat{P}_{\text{hh}}$ or $\hat{P}_{\text{ind}}$ depending on the prediction level.

Balanced Subsampling Strategy

TabPFN's in-context learning mechanism imposes a practical constraint on training set size. Self-attention over the input sequence (3.11) has $\mathcal{O}(N_{\text{tr}}^2)$ memory complexity, where N_{tr} is the number of training examples, and the model's pre-training used datasets of bounded size; supplying a context larger than encountered during pre-training constitutes an out-of-distribution scenario. The practical upper bound is $N_{\text{tr}}^{\max} = 45{,}000$ training examples per inference call.

The full training pools in this study substantially exceed this limit: the household-level dataset contains over 250,000 households, and the individual-level dataset contains approximately 612,000 individuals. A single subsample of at most $N_{\text{tr}}^{\max}$ rows is therefore drawn from the training pool for each cross-validation fold. Both datasets also exhibit class imbalance; the proportion of positive examples (households or individuals with a secondary transmission event) is $\approx 18.8\%$ at the household level and $\approx 9.7\%$ at the individual level under the natural distribution. Training on a heavily imbalanced subsample degrades positive-class discrimination; the **positive-class oversampling ratio** $r \in \{0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70\}$, which controls the fraction of positive examples in the training subsample, is therefore treated as a hyperparameter.

The optimal r is selected via a 5-fold grid search: for each candidate ratio, one TabPFN model is trained on a balanced subsample drawn from each fold's training pool, and evaluated on that fold's validation set using macro-averaged AUC. The mean validation AUC across all five folds determines the optimal r, which is then fixed and applied uniformly to all five folds. For each fold, the final training subsample $\mathcal{D}_{\text{tr}}^{(r)}$ is drawn by stratified sampling without replacement, satisfying:

$$|\mathcal{D}_{\text{tr}}^{(r)}| = N_{\text{tr}}^{\max}, \quad \frac{\sum_k \mathbf{1}[y_k = 1]}{N_{\text{tr}}^{\max}} = r, \tag{3.14}$$

where $\mathbf{1}[\cdot]$ is the indicator function. This procedure is performed independently for the household-level and individual-level TabPFN models, as the optimal r may differ between levels due to their different natural class prevalences ($p_{\text{true}} = 0.188$ and $p_{\text{true}} = 0.097$, respectively).

Prior Correction for Natural-Distribution Evaluation

Training on a balanced subsample with positive rate r preserves ranking (AUC) but shifts probability calibration: predicted probabilities $\hat{p}$ are systematically higher than the true marginal $P(Y = 1)$ in the natural population. When evaluating

on a natural-distribution test set with true positive prevalence p_{true}, a log-odds correction is applied to recover calibrated probabilities:

$$\text{logit}(\hat{p}_{\text{corr}}) = \text{logit}(\hat{p}) + \text{logit}(p_{\text{true}}) - \text{logit}(r), \tag{3.15}$$

where $\text{logit}(p) = \log\big(p/(1-p)\big)$. This correction is applied separately to each TabPFN model using the corresponding p_{true} and r values from that model's training configuration. It preserves the AUC exactly and corrects the probability scale, enabling meaningful threshold-based evaluation at the true class prevalence. The three baseline classifiers (logistic regression, random forest, XGBoost) are trained on the full original training set without subsampling and therefore do not require this correction.

3.3 Explainability Methods

Model transparency is achieved through SHAP (SHapley Additive exPlanations) [19], applied at three levels of granularity to both prediction models. SHAP grounds feature attribution in cooperative game theory [20]: for any sample $\mathbf{x} \in \mathbb{R}^d$, the model prediction is decomposed as

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{d} \phi_j(\mathbf{x}), \tag{3.16}$$

where $\phi_0 = \mathbb{E}_{\mathbf{X}}[f(\mathbf{X})]$ is the global baseline and $\phi_j(\mathbf{x})$ is the *SHAP value* of feature j: its signed contribution to the deviation of the prediction from the baseline. Concretely, ϕ_j is the feature's average marginal contribution across all possible coalitions $S \subseteq \mathcal{F} \setminus \{j\}$:

$$\phi_j(\mathbf{x}) = \sum_{S \subseteq \mathcal{F}\setminus\{j\}} \frac{|S|!\,(d-|S|-1)!}{d!} \Big[v_{\mathbf{x}}(S \cup \{j\}) - v_{\mathbf{x}}(S) \Big], \tag{3.17}$$

where $v_{\mathbf{x}}(S) = \mathbb{E}_{\mathbf{X}_{\bar{S}}}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})]$ is the expected model output when only the features in S are observed and the remaining features $\bar{S} = \mathcal{F} \setminus S$ are marginalised over the data distribution. The additive decomposition (3.16) guarantees that SHAP values sum to the full prediction deviation from the baseline, making them interpretable as fair, signed marginal contributions rather than arbitrary importance scores.

3.3.1 KernelSHAP

Exact computation of (3.17) requires evaluating $v_{\mathbf{x}}$ for all 2^d feature subsets, which is computationally intractable for the feature dimensions in this study. KernelSHAP [19] provides a model-agnostic approximation based on the following key insight: the Shapley values ϕ_j in (3.17) can be recovered as the coefficients of a weighted linear model fitted over coalitions. To see this, each coalition $S \subseteq \mathcal{F}$ is encoded as a binary indicator vector $\mathbf{z} \in \{0,1\}^d$, where $z_j = 1$ if feature j is included in S and $z_j = 0$ otherwise. The value function $v_{\mathbf{x}}(S)$ is then approximated

by a linear surrogate $g(\mathbf{z}) = \phi_0 + \sum_j \phi_j z_j$ whose coefficients are exactly the SHAP values we seek. Absent features ($z_j = 0$) are marginalised by replacing them with values drawn from a background dataset $\mathcal{B} \subset \mathbb{R}^d$ of size $m = |\mathcal{B}|$:

$$\hat{v}_{\mathbf{x}}(\mathbf{z}) = \frac{1}{m} \sum_{l=1}^{m} f\big(\mathbf{x} \odot \mathbf{z} + \mathbf{b}^{(l)} \odot (\mathbf{1} - \mathbf{z})\big), \tag{3.18}$$

where $\mathbf{b}^{(l)} \in \mathcal{B}$ is the l-th background sample used to approximate the marginalisation over absent features, and $\odot$ denotes element-wise multiplication. Intuitively, for positions where $z_j = 1$ the original feature value x_j is retained; for positions where $z_j = 0$ it is replaced by the background value $b_j^{(l)}$, and the results are averaged over the m background samples. The SHAP coefficients $\boldsymbol{\phi}$ are then recovered by minimising a weighted least-squares objective over a random sample of M coalitions,

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \sum_{t=1}^{M} \pi_{\mathbf{x}}(\mathbf{z}^{(t)}) \big[g(\mathbf{z}^{(t)}) - \hat{v}_{\mathbf{x}}(\mathbf{z}^{(t)})\big]^2, \tag{3.19}$$

where the kernel weight $\pi_{\mathbf{x}}(\mathbf{z}) = (d-1) \big/ \Big[\binom{d}{|\mathbf{z}|} |\mathbf{z}| (d - |\mathbf{z}|)\Big]$, with $|\mathbf{z}| = \sum_j z_j$ denoting the coalition size (number of active features), upweights coalitions of extreme size (i.e. very small or very large $|\mathbf{z}|$), ensuring convergence to the true Shapley values as $M \to \infty$.

3.3.2 Three-Level Explanation Structure

XAI analyses at both the household and individual levels are structured in three tiers of increasing specificity, each serving distinct analytical and communicative purposes.

Global SHAP is computed on a *class-balanced explain set*: a fixed subset of the test set containing equal numbers of positive and negative instances, sampled without replacement to ensure that summary statistics are not dominated by the majority class. The results are visualised as a *beeswarm plot*, in which each point represents the SHAP value $\phi_j(\mathbf{x})$ of one instance for a given feature j. Features are ranked vertically by their mean absolute SHAP value $\bar{\phi}_j = \mathbb{E}[|\phi_j(\mathbf{x})|]$ across the explain set, reflecting overall feature importance. Point colour encodes the original (pre-standardisation) feature value (red for high values, blue for low values), so the plot simultaneously reveals which features matter most and whether high or low values are associated with higher predicted risk. Global SHAP analysis serves primarily as a model validation tool to assess whether the learned feature–outcome relationships are epidemiologically plausible.

The test set is then partitioned into three risk subgroups based on the model's predicted probability $\hat{p}$. Rather than applying fixed arbitrary thresholds, the strata are defined relative to the optimal classification threshold τ^* obtained via F1-score maximisation on the validation set: $\tau^*_{\text{ind}} = 0.192$ for the individual-level model and $\tau^*_{\text{hh}} = 0.256$ for the household-level model. Stratum boundaries are set at $\tau^*/2$

and $3\tau^*/2$, yielding:

Individual level:

$$\text{Low: } \hat{p} < 0.096, \quad \text{Moderate: } 0.096 \leq \hat{p} < 0.288, \quad \text{High: } \hat{p} \geq 0.288. \tag{3.20}$$

Household level:

$$\text{Low: } \hat{p} < 0.128, \quad \text{Moderate: } 0.128 \leq \hat{p} < 0.384, \quad \text{High: } \hat{p} \geq 0.384. \tag{3.21}$$

These threshold-derived boundaries ensure that the moderate-risk stratum straddles the operational classification boundary, making the subgroup analysis directly relevant to the model's deployment behaviour. They are used exclusively for SHAP subgroup stratification and do not constitute clinical risk cutoffs. A beeswarm SHAP analysis is performed separately for each subgroup, revealing how the model's decision logic shifts across the risk spectrum and providing epidemiologically relevant evidence about risk-stratified intervention targets.

Finally, local SHAP values are computed for individual instances selected from each risk subgroup, producing *waterfall plots* that decompose a single prediction into signed per-feature contributions starting from the global baseline ϕ_0. Each bar in a waterfall plot represents $\phi_j(\mathbf{x})$ for one feature, and the bars accumulate from ϕ_0 to the final predicted value $f(\mathbf{x})$, making the contribution of each feature to that specific prediction immediately readable. Instances are chosen to be representative of their subgroup's explanation pattern. Concretely, each instance is represented as a vector of its d SHAP values $(\phi_1(\mathbf{x}), \ldots, \phi_d(\mathbf{x}))$; the subgroup centroid is the component-wise mean of these vectors across all subgroup members, and the selected instance is the one whose SHAP vector is closest to this centroid under Euclidean distance. A waterfall plot for a specific household or individual constitutes the structured input passed to the LLM-based explanation layer (Section 3.4), which translates it into natural language for non-technical stakeholders.

3.4 LLM-Based Explanation System

The final component of the system design bridges the technical outputs of the SHAP pipeline and the non-technical stakeholders who must act on them. Given a structured inference report for a single household, containing the ML predictions and ranked SHAP attributions for both levels, a large language model (LLM) is prompted to produce a structured, plain-language interpretation targeted at policy-oriented readers with no ML background. This section describes the design of that LLM layer: Section 3.4.1 specifies the structured inference report that serves as input; Section 3.4.2 describes the prompting strategy; Section 3.4.3 lists the LLMs evaluated; and Section 5.1 introduces the two-tier evaluation framework used to assess explanation quality.

3.4.1 Inference Report Format

The primary input to the LLM explanation layer is a structured plain-text inference report (hereafter `report.txt`) generated automatically by the inference pipeline for each household in the evaluation set. The report is self-contained

and encodes all information the LLM requires without any prior knowledge of the dataset or model architecture. It covers three categories of information. First, the household composition section provides the household identifier, total size, index case member IDs, and non-index member IDs. Second, the household-level model output section contains the predicted probability of secondary transmission $P(\text{transmission})$, the empirical decision threshold, the binary model flag (TRANSMISSION FLAGGED / NO TRANSMISSION FLAGGED), and the top-ranked household SHAP features with their pre-standardisation values and SHAP directions (HIGHER / LOWER). Third, the individual-level model output section provides, for each non-index member, the predicted susceptibility score $P(\text{susceptible})$, the decision threshold, the binary flag, and the top-ranked individual SHAP features with pre-standardisation values and directions.

Crucially, `report.txt` does *not* include ground truth labels, reflecting a realistic prospective deployment scenario in which it is unknown at inference time whether a predicted transmission event will actually occur. The LLM is therefore explicitly prohibited from claiming knowledge of whether any prediction was correct.

3.4.2 Prompt Engineering

The LLM explanation system uses a two-part prompt structure: a **system prompt** fixed across all household cases, and a **chain-of-thought (CoT) user prompt** instantiated per case.

The system prompt defines the LLM's role, scope constraints, and interpretive framework. The LLM is positioned as an expert machine-learning analyst and science communicator whose task is to translate model inference results into plain language for public health policy experts with no ML background. The centrepiece of the system prompt is a critical scope constraint: all outputs must be framed as explanations of *model behaviour*, not as clinical diagnoses, epidemiological ground truth, or verified infection risk estimates. In practice this means the LLM is instructed that probabilities reflect model scores based on patterns in Swedish registry data rather than true infection probabilities, that decision thresholds are empirical ML values derived from F1-score optimisation rather than clinical cutoffs, that SHAP values explain the model's prediction for a specific case without establishing causation, and that ground truth is not available at inference time and must not be assumed or speculated about. The system prompt also describes the two-level prediction architecture: the household model outputs a transmission probability and flag, and the individual model outputs per-member susceptibility scores. The system prompt also explains their complementary roles: the household level determines whether intervention is model-indicated, while the individual level identifies priority members within flagged households. Finally, the system prompt embeds a full feature glossary mapping technical variable names (e.g., `TRYGG_1`, `lmed_R03A`) to their Swedish registry source and plain-language English interpretation, enabling the LLM to produce descriptions such as "homecare alarm activations (a proxy for underlying frailty)" rather than reporting raw variable names.

The CoT user prompt is instantiated once per household and instructs the

LLM to structure its output into five sections addressing distinct communication needs. The *Case Summary* (3–4 sentences) describes the household composition and the model's predictions in plain language, with explicit framing templates ensuring that probabilities are always anchored as model outputs rather than real-world risk (e.g., "the model assigned a transmission score of 73% to this household, above the model's flagging threshold of 50%"). The *Household Model Assessment* interprets the household-level output by reporting the predicted score relative to the threshold, explaining the practical meaning of the most influential household features, and performing a domain-coherence check on whether the overall feature pattern aligns with known epidemiological risk factors. The *Individual Susceptibility Profiles* section covers each non-index member in turn, providing a model-framed score statement, identifying the top 2–3 driving features with their plain-language interpretation, and including a domain-plausibility note without implying whether the model was correct. The *Expert Considerations* section highlights domain-consistent and uncertain patterns in the model's reasoning, calls attention to features acting as proxies for unmeasured factors (e.g., income as a proxy for housing quality; homecare alarms as a proxy for frailty), and flags predictions close to the decision threshold as uncertain. The *Model Confidence Note* (2–3 sentences) assesses the certainty of the model's outputs using threshold-relative framing and closes by reiterating that prediction correctness is unknown and that outputs serve as a limited reference for qualified experts rather than a clinical decision tool. In addition to these structural requirements, the CoT prompt specifies tone and style constraints: the target audience is a senior public health researcher or policy analyst; technical ML terminology must be replaced with plain-language equivalents; concrete language referencing actual feature values is preferred over abstract statistical descriptions; and the target length is 500–700 words.

3.4.3 LLMs Selection

Five frontier LLMs were selected for evaluation, spanning the major proprietary and open-source model families available at the time of the study: Claude Sonnet 4.5 (Anthropic), GPT-5.3 (OpenAI), Grok 4.2 (xAI), DeepSeek V3.2 (DeepSeek), and Llama 4 (Meta). All five models were accessed via their respective public APIs (or, in the case of Llama 4, a self-hosted inference endpoint) with version numbers pinned throughout the evaluation to prevent version-drift confounds.

The selection rationale was deliberately broad rather than narrowly optimised. The primary objective of this evaluation is not to compare the relative capabilities of these models against one another, but to *validate the pipeline itself*: to establish that the structured inference report format, the chain-of-thought prompting strategy, and the two-tier evaluation framework are jointly sufficient to elicit and assess high-quality policy-oriented explanations from current state-of-the-art LLMs. Selecting models from distinct architectural lineages (instruction-tuned proprietary transformers at different scales and training regimes, and a leading open-source alternative) provides coverage of the deployment scenarios most likely to be encountered in practice. Should future applications of this pipeline require adaptation to domain-specific or computationally constrained settings, the evalua-

tion framework developed here can be reapplied directly to assess whether a target model meets the quality thresholds demonstrated by the frontier models evaluated in this study.

Each model was evaluated on ten independent runs using the identical system prompt and CoT user prompt described in Section 3.4.2, applied to the same `report.txt` file for each household case. No fine-tuning or retrieval augmentation was applied; all models were evaluated in a zero-shot prompting regime. Running ten repeated trials per model and case allows the detection of output variance attributable to the LLMs' stochastic generation, and permits mean scores and standard deviations to be reported for each metric rather than relying on a single evaluation pass.

Chapter 4
Experiments and Results

This chapter reports the empirical implementation and evaluation of the system described in Chapter 3. Section 4.1 describes the register data sources, the concrete feature sets constructed for each prediction level, and the preprocessing decisions applied at each level. Section 4.2 specifies the study population, evaluation protocol, and the full hyperparameter configuration used in all experiments. Section 4.3 reports model performance at both prediction levels. Section 4.4 presents and interprets the three-level XAI analysis.

4.1 Data Sources, Feature Construction, and Preprocessing

4.1.1 Swedish National Population and Healthcare Registers

This thesis uses data from Swedish population and healthcare registers accessed through SWECOV (Swedish Register-based Research Project on COVID-19) [41, 42]. The thesis draws on record-linkage across five Swedish national registers covering January–December 2020, corresponding to the first SARS-CoV-2 epidemic wave prior to vaccination. Records are linked at the individual level using the Swedish personal identity number (personnummer). Household membership is operationalized as co-residence at the same registered address on 1 January 2020, derived from the Total Population Register (RTB). Ethical permission was granted by the Swedish Ethical Review Authority (permit numbers 2020-06492, 2021-01115, 2022-01355-02, 2022-06118-02, and 2024-02342-02).

4.1.2 Preprocessing Pipeline

Figure 4.1 provides an overview of the data pipeline connecting the Swedish national population registers to the final cross-validated datasets used for model training and evaluation. The pipeline proceeds in five sequential steps: data cleaning, household-level missing value imputation, medical history filtering, feature engineering, and stratified dataset splitting. The following subsections describe each step in turn.

Step 1: Data Cleaning. Raw register records are merged on the personal identity number and filtered to the study population defined in Section 3.1.1. Im-

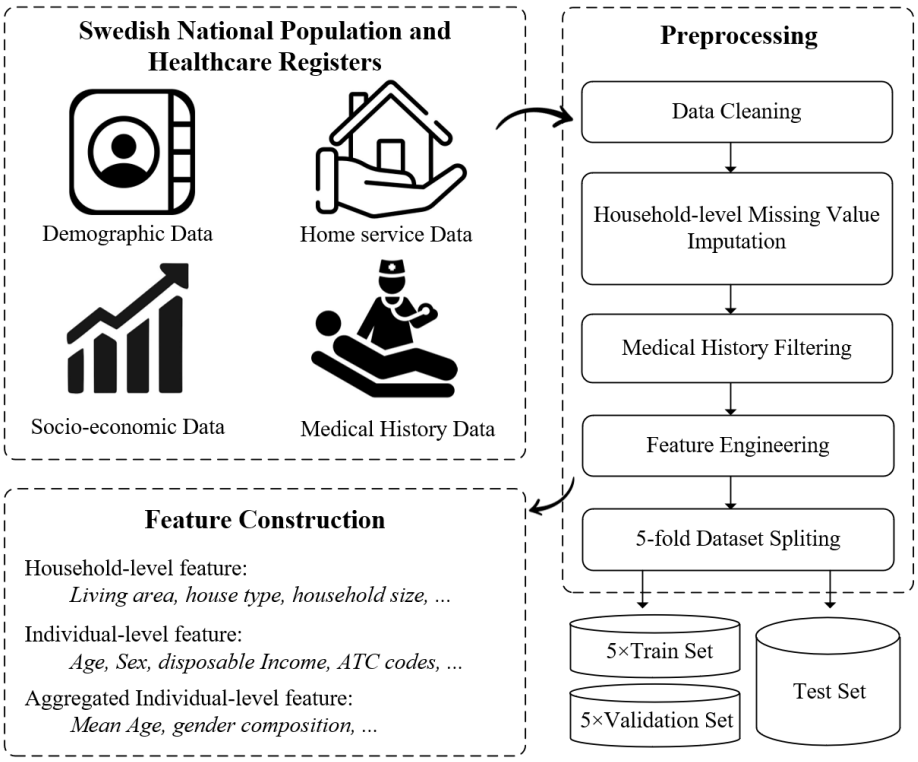


Figure 4.1: Data pipeline from Swedish national population registers to the cross-validated datasets used for model training.

plausible values (e.g. negative income entries, diagnosis dates outside the study window) are treated as missing. Categorical variables are harmonized to a consistent encoding across registers, and the housing type variable is one-hot encoded with unseen categories mapped to an unknown class.

Step 2: Household-level Missing Value Imputation. Following data cleaning, continuous household-level features with a missing rate exceeding 20% in the training fold are dropped entirely. Remaining numeric missing values at the household level are imputed with the training-fold median. This step is applied to the household dataset only; at the individual level, medical history features are binary indicators and therefore require no numeric imputation.

Step 3: Medical History Filtering. Individual-level medical history is initially represented by a large set of binary indicators encoding whether each drug class (ATC code) or diagnosis category (ICD code) appeared in the individual's records over the two years prior to their index date. To keep dimensionality tractable, these indicators pass through a two-stage filtering procedure applied within each training fold. First, a χ^2 association test ($\alpha = 0.05$) removes statisti-

cally uninformative features. Second, a near-zero-variance filter removes features where one category dominates $\geq 95\%$ of observations. Surviving indicators are further restricted to an epidemiological whitelist of drug classes and diagnosis categories with established links to COVID-19 susceptibility or severity, covering respiratory, cardiometabolic, immunological, oncological, and endocrine conditions, as well as immunosuppressant and anti-infective drug use.

Step 4: Feature Engineering. After cleaning, imputation, and filtering, the three feature categories illustrated in Figure 4.1 are constructed.

Household structural features ($\mathbf{c}_j$) are attributes that are constant across all members of household j: household size, per-capita living area, housing type, and similar dwelling-level variables. These are included directly as structural features without aggregation.

Aggregated individual-level features ($\mathbf{a}_j$) are constructed by aggregating the demographic, socioeconomic, and homecare attributes of all n_j household members: continuous variables are summarised by their mean, maximum, minimum, variance, interquartile range, and range; binary attributes are represented as proportions or indicator variables (e.g. presence of elderly members, multigenerational structure). An additional group of *index-case features* is computed exclusively from the index cases, capturing their age distribution, sex composition, immigration background, income level, living area, and homecare usage, together with the ratio of index cases to total household size.

Together, $\mathbf{c}_j$ and $\mathbf{a}_j$ form the household-level feature vector $\mathbf{x}_j^{\text{hh}} = [\mathbf{a}_j;\, \mathbf{c}_j]$ used by the household-level model.

Individual-level features ($\mathbf{u}_i$, forming $\mathbf{x}_i^{\text{ind}} = [\mathbf{u}_i;\, \mathbf{a}_{j(i)};\, \mathbf{c}_{j(i)}]$) consist of the individual's own demographic and socioeconomic attributes and the filtered medical history indicators. The aggregated features $\mathbf{a}_{j(i)}$ and structural features $\mathbf{c}_{j(i)}$ of the individual's household are appended to provide household context, making them shared components between the two models.

Financial and area variables are log-transformed to reduce skew, and all continuous features are standardised to zero mean and unit variance, excluding binary, count, and categorical indicator columns.

Step 5: Dataset Splitting. The full eligible population is divided into a stratified 10% hold-out test set drawn first, and the remaining 90% is split into five stratified folds for cross-validation. This yields five train/validation fold pairs (5×Train Set and 5×Validation Set) plus a single held-out Test Set, as shown in Figure 4.1. Stratification is performed on the binary outcome label at each prediction level to preserve the natural class ratio across all splits. All transformations are fitted exclusively on the training fold of each cross-validation split and applied without refit to the validation and test sets, preventing any leakage of test-set statistics into the training procedure.

Table 4.1: Descriptive statistics of the household-level and individual-level datasets.

Level	Characteristic	Value
Household	Total households	252,472
	Transmission occurred (positive)	47,344 (18.8%)
	No transmission (negative)	205,128 (81.2%)
	Mean household size	3.52
Individual	Total individuals	608,473
	Secondary case (positive)	58,954 (9.7%)
	Not infected (negative)	549,519 (90.3%)
	Mean age in 2020, years	31.8
	Female	47.9%
	Foreign background	22.7%
	Mean disposable income (100 SEK/year)	1,892
	Receiving home care	0.1%

4.2 Experimental Configuration

4.2.1 Study Population

Table 4.1 summarises the characteristics of the two study populations after applying the eligibility criteria from Section 3.1.1.

4.2.2 Positive-Class Ratio Grid Search

As established in Section 3.2.5, the positive-class oversampling ratio r controls the fraction of positive examples in each training subsample and is the primary hyperparameter governing the precision–recall trade-off of the TabPFN models. Because r directly shapes the model's implicit prior over class probabilities, its selection is treated as a formal hyperparameter search rather than a fixed choice.

For both the household-level and individual-level models, the search is conducted across all five cross-validation folds. For each candidate value of r, one TabPFN model is trained on a balanced subsample of size $N_{\text{tr}}^{\max} = 45{,}000$ drawn from each fold's training set, and evaluated on that fold's validation set using macro AUC as the selection criterion. The mean validation AUC across the five folds determines the optimal r, which is then fixed and applied uniformly to all five final training runs.

The search grid spans $r \in \{0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70\}$. Table 4.2 reports the mean macro AUC across the five folds for each candidate value at both prediction levels.

The grid search identifies $r = 0.60$ as the optimal ratio for the household-level model and $r = 0.50$ for the individual-level model. The household model benefits from a more aggressive oversampling of the positive class (natural prevalence

Table 4.2: TabPFN positive-class ratio grid search results (mean validation AUC-ROC across 5 folds ± std). The optimal ratio per level is highlighted in bold; the selection criterion is mean validation macro AUC-ROC.

Positive ratio	Individual level AUC-ROC	Household level AUC-ROC
0.10	0.7125 ± 0.0027	0.6270 ± 0.0015
0.20	0.7181 ± 0.0028	0.6363 ± 0.0017
0.30	0.7203 ± 0.0030	0.6391 ± 0.0016
0.40	0.7213 ± 0.0028	0.6403 ± 0.0019
0.50	**0.7215 ± 0.0027**	0.6409 ± 0.0022
0.60	0.7214 ± 0.0024	**0.6411 ± 0.0022**
0.70	0.7207 ± 0.0021	0.6406 ± 0.0023

18.8%), while the individual model, with a more severe class imbalance (natural prevalence 9.7%), achieves its best validation AUC at a more moderate 50% positive fraction. In both cases, the selected ratio substantially overrepresents the minority class relative to its natural prevalence, pushing the model's implicit prior toward positive predictions and improving recall without collapsing precision to an unacceptable level. The prior correction defined in Equation (3.15) subsequently restores calibrated probabilities when the models are evaluated on the original test sets at natural class prevalence.

4.2.3 Experimental Setup

All experiments were conducted on Bianca, the sensitive-data high-performance computing cluster operated by UPPMAX (Uppsala Multidisciplinary Center for Advanced Computational Science) at Uppsala University, provisioned through the National Academic Infrastructure for Supercomputing in Sweden (NAISS) [43]. Bianca is the designated national resource for research involving sensitive personal data, providing an isolated, air-gapped environment with no direct internet access and encrypted, logged data transfers, which satisfies the data governance requirements of the Swedish register linkage used in this study.

TabPFN training and inference were executed on GPU nodes equipped with an NVIDIA A100-PCIE-40GB graphics processor (40 GB VRAM, Ampere architecture, 108 streaming multiprocessors), 16 CPU cores, and 256 GB RAM. The GPU is required for the transformer self-attention mechanism in TabPFN, which would be prohibitively slow on CPU for the subsample sizes used here. All baseline classifiers (logistic regression, random forest, and XGBoost) and KernelSHAP computations were executed on CPU nodes. The standard Bianca CPU nodes provide dual Intel Xeon E5-2630 v3 processors (8 cores per socket, 16 cores total) with 128 GB RAM, while the newer Maja nodes provide an AMD EPYC 9454P processor (48 cores) with 768 GB RAM. All code was executed under Python 3.12.12.

Table 4.3 lists the hyperparameters for all four classifiers and the KernelSHAP explainer, each referenced to its corresponding symbol in the methodology. Logistic regression and random forest address class imbalance through inverse-frequency sample weights (w_+, w_-); XGBoost uses the equivalent `scale_pos_weight` mech-

Table 4.3: Hyperparameter configuration for all models and the KernelSHAP explainer, mapped to their methodological symbols. Rows for parameters without a defined symbol in Chapter 3 are omitted.

Model	Parameter (symbol)	Value
Logistic Regression	Regularisation strength λ_{LR} (3.5)	1.0
	Sample weights $w_+ = N/(2N_+),\ w_- = N/(2N_-)$	inverse-frequency
Random Forest	Ensemble size B (3.6)	200
	Sample weights $w_+ = N/(2N_+),\ w_- = N/(2N_-)$	inverse-frequency
XGBoost	Learning rate η (3.8)	0.05
	Minimum loss reduction γ (3.10)	0
	Leaf-score regularisation λ_{XGB} (3.10)	1
	Positive-class weight N_-/N_+ (`scale_pos_weight`)	fold-specific
TabPFN	Training subsample cap $N_{\text{tr}}^{\max}$ (3.14)	45,000
	Positive-class ratio r — household (3.14)	0.60
	Positive-class ratio r — individual (3.14)	0.50
KernelSHAP	Background size $m = \lvert\mathcal{B}\rvert$ (3.18)	50
	Coalition samples M — global, household (3.19)	200
	Coalition samples M — global, individual (3.19)	300
	Coalition samples M — subgroup / local (3.19)	100

anism set to N_-/N_+ per fold. TabPFN is trained on a balanced subsample of $N_{\text{tr}}^{\max} = 45{,}000$ rows with positive-class ratio r selected by the 5-fold grid search described in Section 4.2.2.

4.2.4 Evaluation Protocol and Metrics

All models are evaluated on the original hold-out test set at the natural class prevalence (18.8% positive at the household level; 9.7% at the individual level). For TabPFN, predicted probabilities are prior-corrected via Equation (3.15) to adjust for the imbalance introduced by the balanced training subsample. Per-fold classification thresholds are derived by maximising macro F1 on the original validation set rather than fixed at 0.5.

Four metrics are reported for each model and evaluation protocol. Macro AUC-ROC is threshold-independent and serve as the primary discriminative measure. Positive-class F1 (F1_+) capture sensitivity to the minority class. Macro F1 summarises overall classification quality across both classes. The Matthews Correlation Coefficient (MCC) provides a holistic single-number summary that is robust to class imbalance and accounts for all cells of the confusion matrix.

Table 4.4: Household-level model performance on the hold-out test set (natural prevalence: 18.8% positive), mean ± std across 5 folds. Best result in **bold**; TabPFN probabilities are prior-corrected.

Model	Macro AUC	Macro F1	MCC	Class 1 F1
Logistic Regression	0.6174 ± 0.0004	0.5087 ± 0.0015	0.1381 ± 0.0027	0.3509 ± 0.0015
Random Forest	0.6383 ± 0.0007	0.5550 ± 0.0009	0.1435 ± 0.0015	0.3413 ± 0.0009
XGBoost	0.6412 ± 0.0010	0.4943 ± 0.0098	$\mathbf{0.1578} \pm 0.0017$	$\mathbf{0.3627} \pm 0.0008$
TabPFN	$\mathbf{0.6453} \pm 0.0011$	$\mathbf{0.5683} \pm 0.0011$	0.1422 ± 0.0064	0.3200 ± 0.0113

Table 4.5: Individual-level model performance on the hold-out test set (natural prevalence: 9.7% positive), mean ± std across 5 folds. Best result in **bold**; TabPFN probabilities are prior-corrected.

Model	Macro AUC	Macro F1	MCC	Class 1 F1
Logistic Regression	0.6888 ± 0.0001	0.4844 ± 0.0007	0.1624 ± 0.0005	0.2474 ± 0.0004
Random Forest	0.7217 ± 0.0006	0.5648 ± 0.0010	0.1988 ± 0.0013	0.2856 ± 0.0010
XGBoost	$\mathbf{0.7258} \pm 0.0004$	0.5812 ± 0.0038	$\mathbf{0.2042} \pm 0.0010$	$\mathbf{0.2927} \pm 0.0012$
TabPFN	0.7165 ± 0.0008	$\mathbf{0.5898} \pm 0.0006$	0.1890 ± 0.0030	0.2786 ± 0.0037

4.3 Model Performance

4.3.1 Household-Level Results

Table 4.4 reports 5-fold cross-validated performance for all models on the household-level task. The natural positive-class prevalence is 18.8%.

TabPFN achieves the highest macro AUC-ROC of 0.6453, ahead of XGBoost (0.6412), Random Forest (0.6383), and Logistic Regression (0.6174), consistent with the expectation that its causally pre-trained in-context learning mechanism provides strong rank-order discrimination for the household transmission task. TabPFN also achieves the highest Macro F1 (0.5683), demonstrating competitive threshold-dependent performance with a balanced training ratio of $r = 0.60$. On MCC (0.1422) and Class-1 F1 (0.3200), XGBoost performs slightly better, as its inverse-frequency weighting places greater emphasis on the positive class.

4.3.2 Individual-Level Results

Table 4.5 reports 5-fold cross-validated performance for all models on the individual-level susceptibility task. The natural positive-class prevalence is 9.7%, approximately half that of the household level.

At the individual level, the results reveal a nuanced competitive picture. XGBoost achieves the highest AUC-ROC (0.7258), closely followed by Random Forest

(0.7217), while TabPFN (0.7165) ranks third on this threshold-independent metric. TabPFN achieves the highest Macro F1 (0.5898), indicating that its balanced training distribution yields better aggregate performance across both classes at natural prevalence. XGBoost achieves slightly higher MCC (0.2042 vs. 0.1890) and Class-1 F1 (0.2927 vs. 0.2786), reflecting stronger raw positive-class recall under a threshold-optimised regime. The logistic regression baseline lags on all metrics, consistent with the strong discriminative capacity expected from its transformer-based in-context learning mechanism.

Notwithstanding XGBoost's marginal advantage on certain individual-level metrics, TabPFN is selected as the model for all subsequent XAI analysis on both levels. This choice is primarily driven by its strong discriminative performance and its broader utility within the analytical framework. Specifically, TabPFN is pre-trained on datasets generated from structural causal models (SCMs), encoding inductive biases aligned with real-world data-generating processes. While the SHAP analyses reported here deliver associative attributions rather than causal explanations, this causal pretraining prior positions TabPFN as a natural foundation for future counterfactual and causal inference extensions beyond the scope of the current work.

4.4 XAI Analysis

To manage the computational cost of TabPFN inference, which scales with the number of test samples processed in each forward pass due to the in-context attention mechanism, the xAI analysis operates on a subsample of 5,000 test-set observations at natural class prevalence rather than the full hold-out set. The three-level XAI analysis (global SHAP, subgroup SHAP, local SHAP) follows the structure and the instance selection procedure defined in Section 3.3.2. All analyses use the held-out test set with the TabPFN model from Fold 1. The balanced explain set is drawn independently of the performance evaluation samples reported in Section 4.3.

4.4.1 Global SHAP Analysis

Household Level

The global SHAP beeswarm for the household-level model (Figure 4.2) reveals a feature importance structure dominated by index-case characteristics and structural household factors, providing strong epidemiological validation of the model's learned representations.

The most influential feature is the maximum age among the household's confirmed index cases. Households with older index cases (shown as red dots in the beeswarm) cluster systematically on the positive SHAP side, receiving substantially elevated transmission probability predictions. This is epidemiologically well-grounded: older SARS-CoV-2-infected individuals typically experience more severe and prolonged illness, are associated with higher and more sustained viral shedding, and spend more time at home during isolation, all of which intensify the within-household exposure experienced by susceptible members.

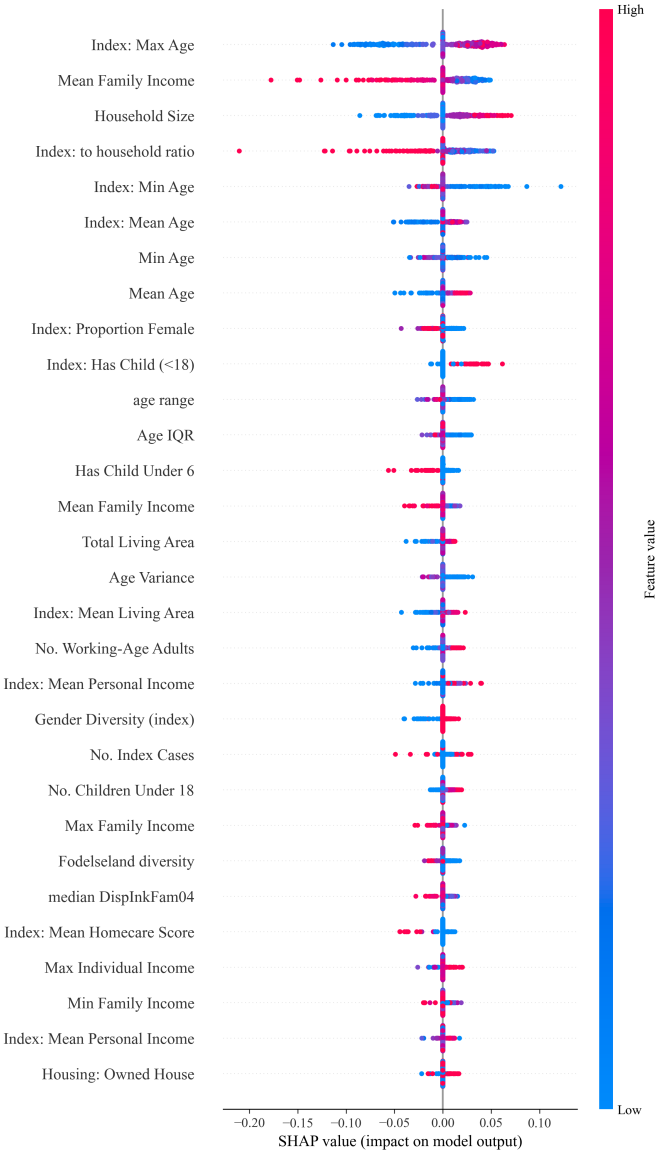


Figure 4.2: Global SHAP beeswarm plot for the household-level TabPFN model (400 balanced samples, features ranked by mean $|\phi_j|$; red = high feature value, blue = low).

The second-ranked feature is the household mean family disposable income. Lower-income households (blue) receive positive SHAP contributions, establishing economic deprivation as a leading independent predictor of household transmission risk. This socioeconomic gradient operates through multiple pathways: lower-income households are more likely to be overcrowded, have limited space

for isolating an infected member, experience greater job insecurity reducing compliance with isolation recommendations, and have restricted access to protective equipment and healthcare.

The third feature, household size, exhibits a clear monotone pattern: larger households (red) receive positive SHAP contributions and smaller households (blue) receive negative ones. This reflects the fundamental contact-opportunity mechanism: more susceptible members imply more exposure events and a higher probability that at least one secondary infection occurs, one of the most consistently replicated findings in household secondary attack rate research [2, 3].

The fourth feature, the ratio of index cases to household size, produces mixed-direction SHAP values, reflecting the interplay between increased transmission pressure from multiple simultaneous infectors and a reduced pool of susceptible members. Further down the ranking, the minimum and mean index-case age, the overall household mean and minimum member age, and age-dispersion measures reinforce the centrality of age structure. The presence of children under six and related child-count variables contribute positively at lower ranks, consistent with children's role as within-household transmission vectors during the pre-vaccination era. Notably absent from the top quartile of the ranking are any clinical or medication-related variables, confirming that at the household level, structural sociodemographic features dominate the transmission signal over individual medical histories.

Individual Level

The global SHAP beeswarm for the individual-level model (Figure 4.3) reveals both continuity with and important divergence from the household-level ranking, reflecting the additional individual-specific predictive signal captured at this finer level of analysis.

The dominant feature at the individual level is family disposable income. Individuals with lower family income (blue) receive systematically positive SHAP contributions, identifying economic vulnerability as the single most important predictor of individual susceptibility. This individual income effect is conceptually distinct from the household mean income effect observed at the household level: whereas the household-level signal reflects the structural economic environment of the household as a whole, the individual income feature captures each member's personal access to economic resources, serving as a proxy for their material ability to protect themselves during household exposure and to seek timely medical attention.

The second feature, the member's own age (Age, 2020), shows a clear positive relationship: older members (red) receive elevated SHAP values, reproducing one of the most robust findings in COVID-19 epidemiology. The third feature, the household maximum index-case age, imports the exposure-intensity signal identified at the household level: the age of the infecting source constitutes an important contextual modifier of individual susceptibility. The fourth feature, the index-to-household ratio, provides further household exposure context.

The number of children under 18 in the household appears sixth with positive contributions, suggesting environments with many children elevate individual sus-

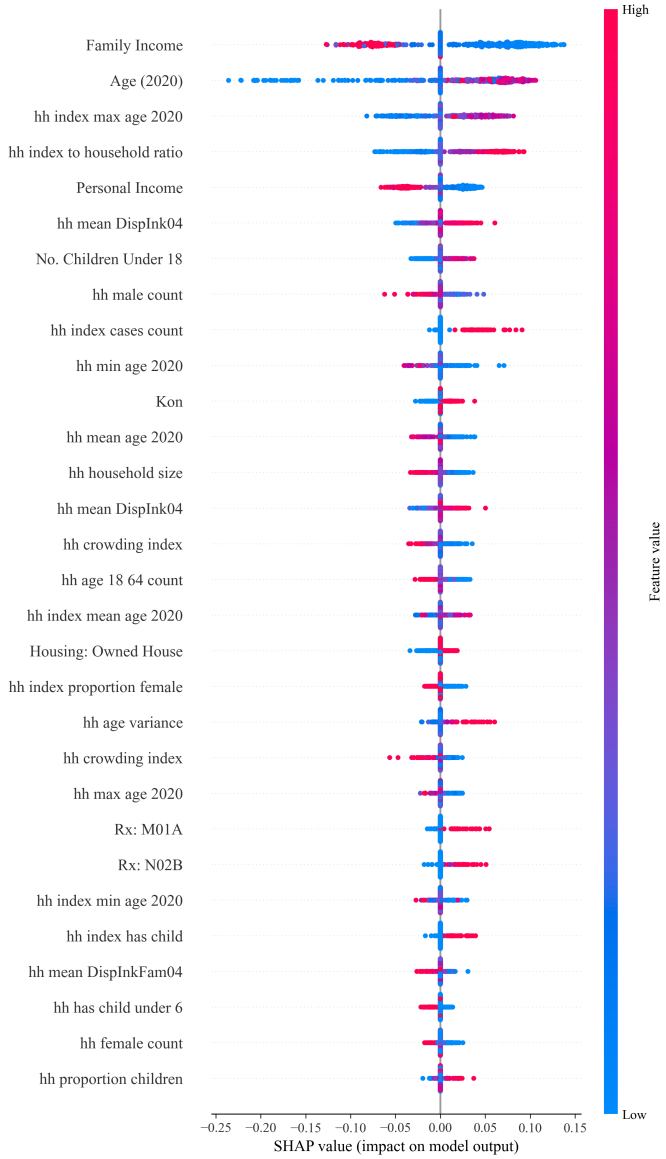


Figure 4.3: Global SHAP beeswarm plot for the individual-level TabPFN model (500 balanced samples, features ranked by mean $|\phi_j|$; red = high feature value, blue = low).

ceptibility through elevated aerosol burden, reduced hygiene control, and higher member rates within shared spaces. Contextual household features, including household mean disposable income, crowding index, age structure, and gender composition, populate the middle ranks, demonstrating the individual model's capacity to incorporate the broader household environment as modifiers of per-

Table 4.6: Subgroup sizes for individual-level and household-level analyses.

Level	Subgroup	N	Proportion
Individual	High risk	3,837	76.7%
	Moderate risk	771	15.4%
	Low risk	392	7.8%
Household	High risk	47	0.9%
	Moderate risk	3,692	73.8%
	Low risk	1,261	25.2%

sonal susceptibility. Medical history variables, including prescriptions for musculoskeletal anti-inflammatory drugs (Rx: M01A, a proxy for chronic inflammatory or rheumatic conditions) and analgesics (Rx: N02B), appear in the lower portion of the top-30 ranking. While their mean absolute SHAP contributions are modest globally, their presence confirms that the model appropriately captures the modulating effect of chronic medical conditions on individual susceptibility, a signal concentrated in specific clinical subgroups and examined further in the subgroup analysis below.

Taken together, the two global SHAP analyses present a coherent and epidemiologically interpretable account of SARS-CoV-2 household transmission risk. At the household level, the age of the index case and socioeconomic deprivation dominate, with household structural characteristics providing complementary signal; at the individual level, the member's own income and age become primary, with household exposure-intensity features as important contextual modifiers. The consistent appearance of income-related features at both levels identifies socioeconomic disadvantage as a cross-cutting driver, operating through mechanisms that span household structure (overcrowding, limited isolation space) and individual vulnerability (access to protection and healthcare).

4.4.2 Subgroup SHAP Analysis

Table 4.6 reports the size of each risk stratum (defined in Eqs. (3.20)–(3.21)) within the explain set at both prediction levels. Risk stratum boundaries are derived from the validation-set optimal classification thresholds: $\tau^*_{\text{ind}} = 0.192$ and $\tau^*_{\text{hh}} = 0.256$.

The distribution reflects the underlying class imbalance at each level. At the individual level, the majority of members fall in the high-risk stratum ($\hat{p} \geq 0.288$, 76.7%), consistent with the relatively high individual-level positive rate in the study population. At the household level, the moderate-risk stratum ($0.128 \leq \hat{p} < 0.384$) dominates at 73.8%, with only 47 households (0.9%) reaching the high-risk threshold ($\hat{p} \geq 0.384$), reflecting the naturally low household-level transmission prevalence (18.8%) and the compressed predicted-probability range of the household model. Beeswarm analyses are presented for all six strata.

Household-Level Subgroup Analysis

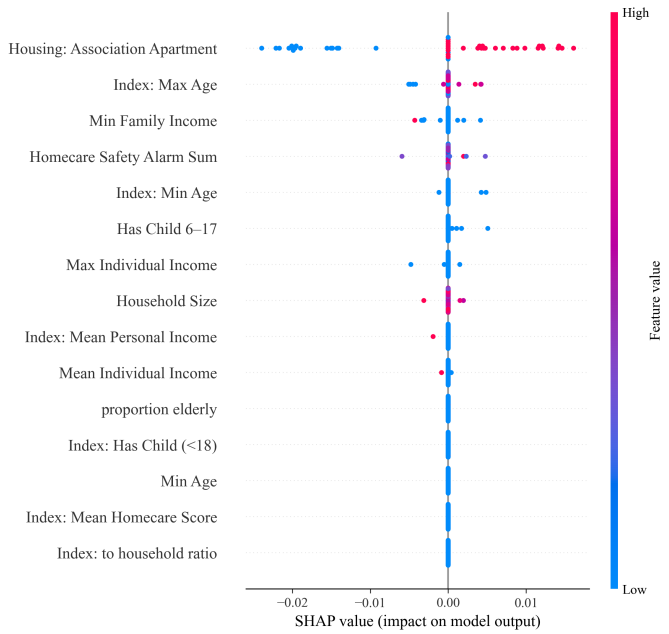


Figure 4.4: Subgroup SHAP beeswarm for the household-level model: high-risk stratum ($\hat{p} \geq 0.384$, $n = 47$ samples). Features are ranked by mean $|\phi_j|$ within the subgroup.

In the household-level high-risk stratum (Figure 4.4, $n = 47$), the leading feature is Housing: Association Apartment (Swedish *bostadsrätt*, a cooperative tenure type common in urban settings). The wide bidirectional spread of SHAP values for this binary indicator reflects that housing tenure is a decisive contextual signal for the highest-risk households, likely proxying for spatial configuration, urban density, and shared-ventilation characteristics associated with multi-unit apartment dwelling. Index: Max Age ranks second, retaining the global finding that older index cases elevate predicted transmission probability. Min Family Income and Homecare Safety Alarm Sum, the latter a proxy for frailty-related care needs among household members, also appear in the top tier, indicating that economic vulnerability and healthcare dependency contribute to the highest-risk household profiles. The overall SHAP magnitudes are substantially smaller (± 0.02) than in the global analysis, reflecting the small stratum size and the compressed probability range near the upper boundary of the score distribution.

In the household-level moderate-risk stratum (Figure 4.5), the leading feature is Födelseland (birth-country) diversity, exhibiting wide spread in both directions. This heterogeneous pattern near the classification boundary suggests that ethnic composition interacts with other household characteristics in a non-monotone manner: it may proxy for unmeasured socioeconomic and behavioural heterogeneity relevant to isolation compliance and within-household exposure management.

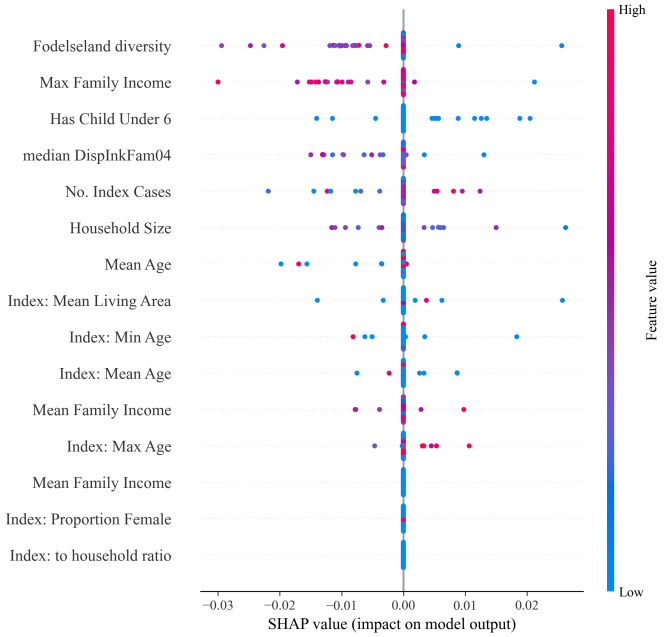


Figure 4.5: Subgroup SHAP beeswarm for the household-level model: moderate-risk stratum ($0.128 \leq \hat{p} < 0.384$, $n = 100$ stratified samples). Features are ranked by mean $|\phi_j|$ within the subgroup.

Max Family Income ranks second, with higher-income households (red) receiving negative SHAP contributions, confirming that economic advantage is protective at the threshold. The presence of children under six (Has Child Under 6), the number of index cases, and household size follow in order, reflecting the combined contribution of exposure intensity and member opportunity in this borderline group.

In the household-level low-risk stratum (Figure 4.6), Födelseland diversity again leads, consistent with the moderate-risk finding and suggesting that ethnic composition is a persistent discriminating variable across the lower portion of the risk spectrum. Mean Age and Index: Max Age rank second and third, with older households and those with older index cases receiving predominantly positive SHAP contributions even within this low-risk group, indicating residual upward age-related risk pressure that keeps certain households approaching the moderate boundary despite their overall low predicted probability. The proportion of children, mean family income, and housing tenure (owned house, which exerts a protective negative contribution) occupy the mid-ranking positions, sketching the profile of a low-risk household: younger, owner-occupied, smaller, and with fewer children.

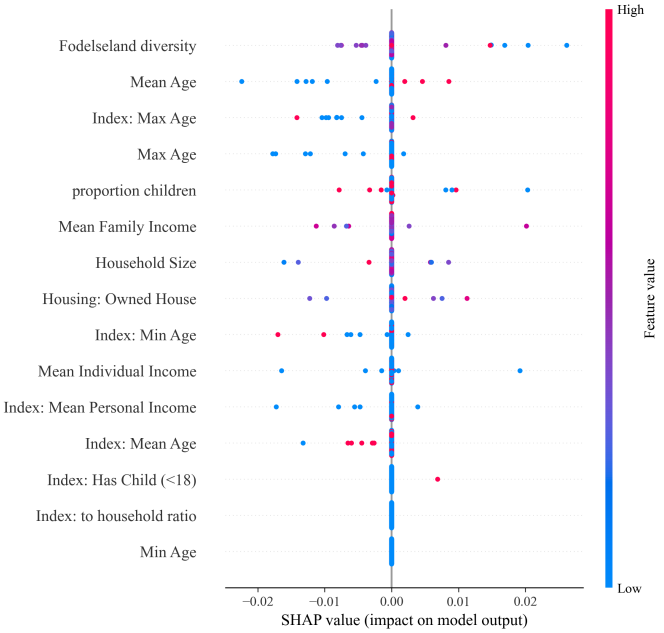


Figure 4.6: Subgroup SHAP beeswarm for the household-level model: low-risk stratum ($\hat{p} < 0.128$, $n = 100$ stratified samples). Features are ranked by mean $|\phi_j|$ within the subgroup.

Individual-Level Subgroup Analysis

In the individual-level high-risk stratum (Figure 4.7), the household crowding index emerges as the dominant feature, with SHAP magnitudes reaching ± 0.15, substantially larger than those observed in the global analysis. Higher crowding values (red) cluster strongly on the positive side while lower crowding (blue) drives predictions toward zero, confirming that household spatial density is especially powerful in distinguishing the most susceptible individuals. This shift from family income (global top feature) to crowding index within the high-risk stratum suggests that, among already-vulnerable individuals, the physical transmission environment becomes the critical differentiating factor. The household mean DispInk04 (disposable income at household level) ranks second with wide spread, reinforcing the socioeconomic signal through a complementary income channel. Family income, household index mean age, and the number of children under 18 follow, maintaining directional consistency with global patterns. A prescription for bronchodilator medication (Rx: R03A, proxying for obstructive lung disease) begins to emerge in the lower portion of the ranking, providing evidence that clinical comorbidities are disproportionately concentrated among the most susceptible individuals, consistent with the known burden of respiratory comorbidities in COVID-19.

In the individual-level moderate-risk stratum (Figure 4.8, $n = 97$), the household mean disposable income and the family income lead with the same pattern

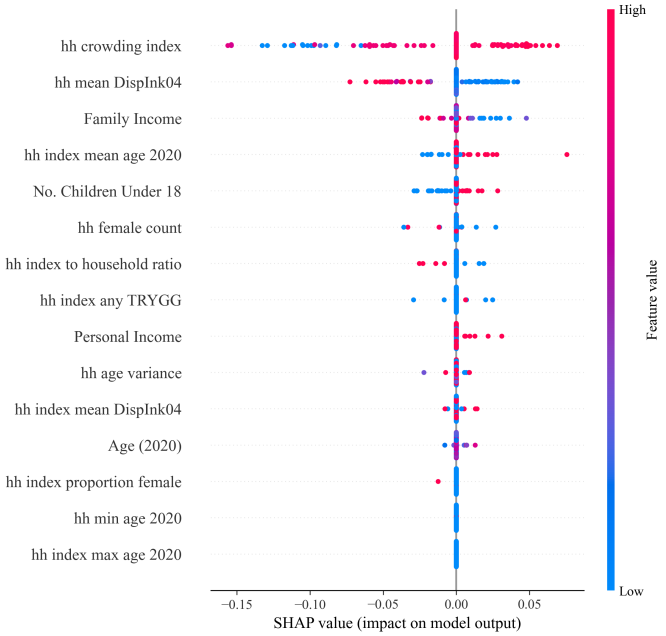


Figure 4.7: Subgroup SHAP beeswarm for the individual-level model: high-risk stratum ($\hat{p} \geq 0.288$, $n = 100$ stratified samples). Features are ranked by mean $|\phi_j|$ within the subgroup.

as global analysis: higher values (red) are associated with negative SHAP contributions. The number of children under 18 and household index-case age features (minimum and mean) rank third and fourth, reflecting the joint contribution of exposure intensity and household composition. Member age occupies sixth position with mixed-direction contributions, indicating that within the moderate stratum individual age is a secondary modifier relative to socioeconomic and household-exposure factors.

In the individual-level low-risk stratum (Figure 4.9, $n = 45$), the household crowding index again leads, now with a strongly left-skewed distribution: the majority of dots cluster at near-zero or negative SHAP values (low crowding = protective), with a small number of high-crowding outliers exerting large positive contributions (up to +0.15). This confirms that very low household crowding is the primary mechanism driving individuals to the low-risk stratum: when a member lives in a spacious household, predicted susceptibility is sharply reduced. Family income and household income measures rank second and third, providing smaller but consistently protective signals. Living area also appears, confirming that spaciousness, measured directly rather than through the crowding index, provides complementary protection. The small stratum size ($n = 45$) limits the precision of individual feature estimates, but the overall pattern is coherent: low-risk individuals live in uncrowded, income-advantaged households with limited transmission pressure.

Across strata, the subgroup analysis corroborates and extends the global SHAP

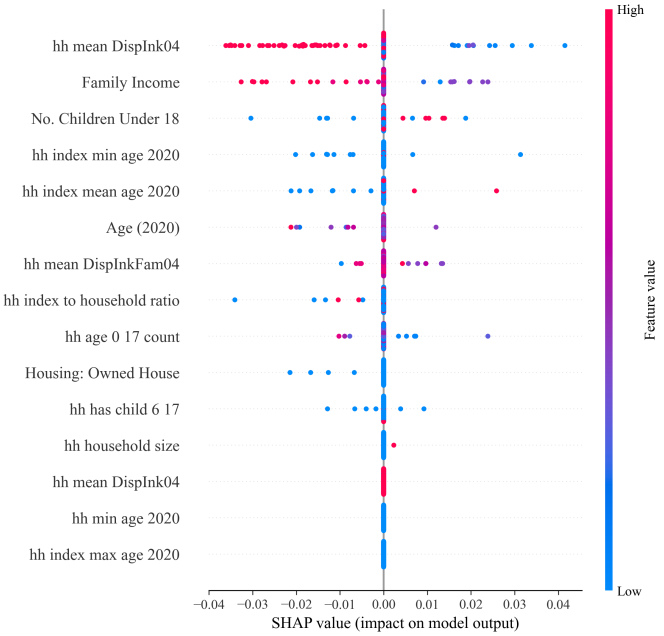


Figure 4.8: Subgroup SHAP beeswarm for the individual-level model: moderate-risk stratum ($0.096 \leq \hat{p} < 0.288$, $n = 97$ stratified samples). Features are ranked by mean $|\phi_j|$ within the subgroup.

findings. At the household level, Födelseland diversity and income features dominate the moderate-to-low risk strata, while housing tenure type distinguishes the rare high-risk households. At the individual level, the crowding index supplants family income as the dominant factor within the high- and low-risk strata, while income-based features remain primary near the classification boundary (moderate stratum). This stratified pattern is consistent with the epidemiological expectation that physical transmission environments matter most at the extremes of the risk distribution, while socioeconomic factors exert their strongest discriminating power near the decision boundary. Together, the household- and individual-level subgroup analyses provide complementary evidence: densely housed, socioeconomically deprived members in households with older index cases represent the most critical target group for preventive intervention.

4.4.3 Local SHAP Analysis

For each risk stratum, one representative instance is selected by proximity to the stratum centroid in SHAP space (see Section 3.3.2). The resulting waterfall plots decompose individual predictions into signed per-feature contributions from the global baseline $E[f(X)]$, and constitute the structured SHAP input passed to the LLM-based explanation layer in Chapter 5.

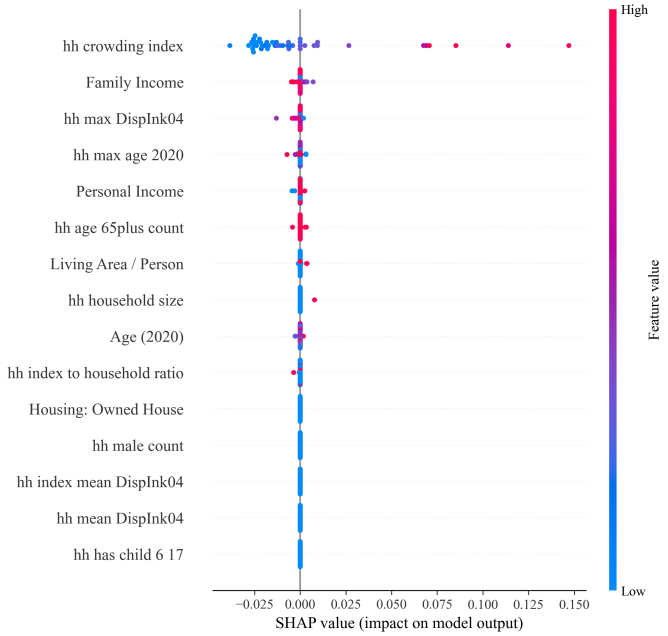


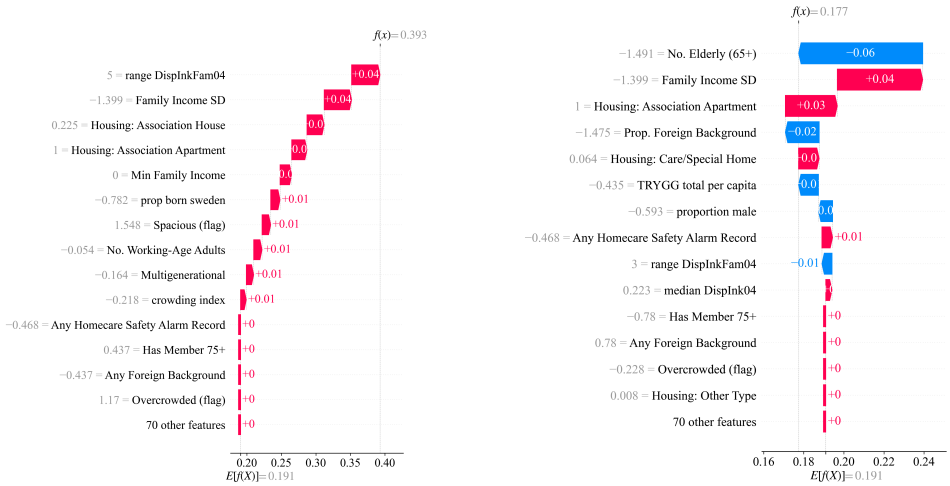
Figure 4.9: Subgroup SHAP beeswarm for the individual-level model: low-risk stratum ($\hat{p} < 0.096$, $n = 45$ stratified samples). Features are ranked by mean $|\phi_j|$ within the subgroup.

Household Cases

Across the three household strata, the waterfall plots reveal several notable patterns. The high-risk case ($\hat{p} = 0.393$) is driven entirely by risk-amplifying features with no competing protective contributions: income dispersion (range DispInkFam04, $+0.04$) and family income standard deviation ($+0.04$) jointly dominate, pointing to a household characterised by marked economic heterogeneity among its members. The moderate-risk case ($\hat{p} = 0.177$) presents a classic competing-risk structure: the number of elderly members (65+) exerts the largest single contribution (-0.06), but is partially offset by family income SD ($+0.04$) and association apartment tenure ($+0.03$), resulting in a net prediction just below the baseline. The low-risk case ($\hat{p} = 0.089$) is driven by a concentrated protective signal from two features alone: the number of elderly members (-0.08) and the proportion of foreign-background members (-0.04), with all remaining features contributing negligibly.

Individual Cases

Across the three individual strata, the household crowding index emerges as the decisive feature at the extremes of the susceptibility distribution. In the high-risk case ($\hat{p} = 0.724$), two crowding index variables jointly contribute $+0.13$, and birth country (frequency-encoded) adds a further $+0.06$; every other feature contributes



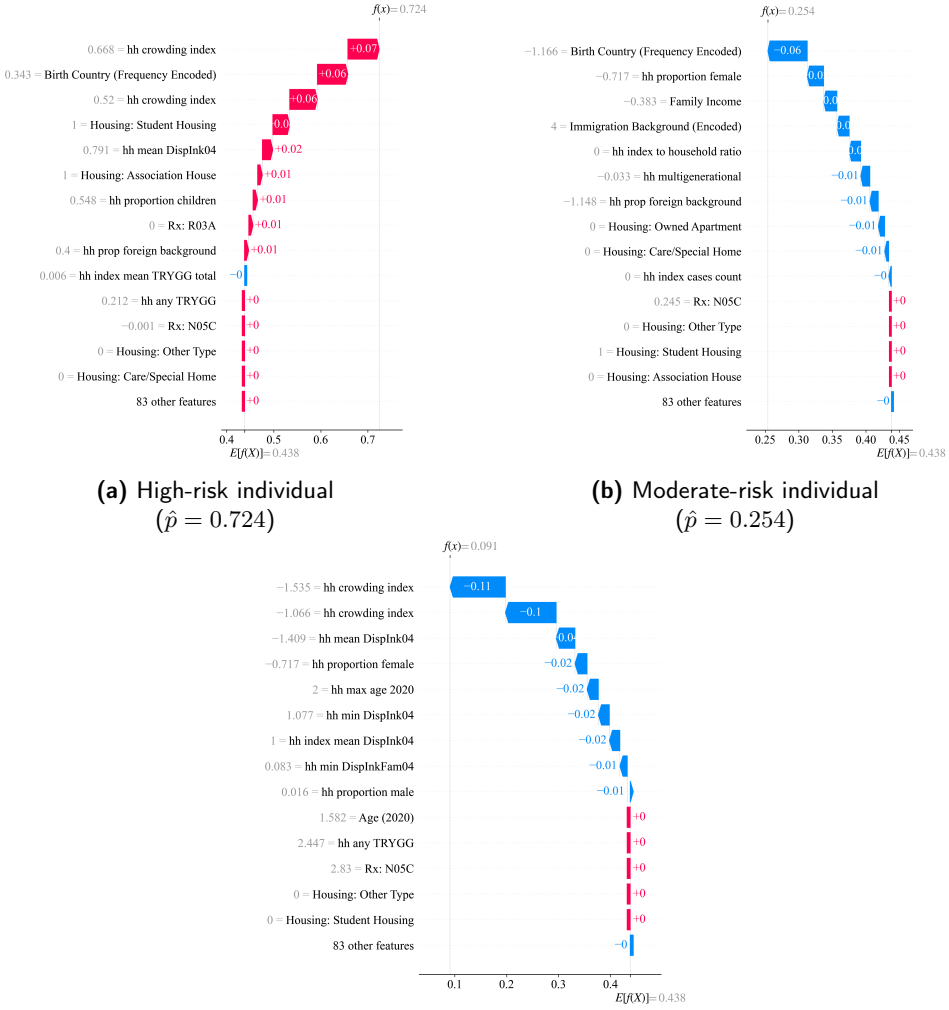
(a) High-risk household ($\hat{p} = 0.393$)

(b) Moderate-risk household ($\hat{p} = 0.177$)

(c) Low-risk household ($\hat{p} = 0.089$)

Figure 4.10: Local SHAP waterfall plots for representative households across all three risk strata (baseline $E[f(X)] = 0.191$). Bars show signed feature contributions ϕ_j from the baseline to the final predicted transmission probability.

positively or is negligible, with no protective offset. In the moderate-risk case ($\hat{p} = 0.254$), birth country reverses direction entirely (-0.06), becoming the dominant contributor, and nearly all remaining features also turn protective, including household proportion female, family income, and immigration background encoding. This directional reversal of birth country across strata reflects the sensitivity of the frequency encoding to population-specific exposure patterns in the register.



(a) High-risk individual ($\hat{p} = 0.724$)

(b) Moderate-risk individual ($\hat{p} = 0.254$)

(c) Low-risk individual ($\hat{p} = 0.091$)

Figure 4.11: Local SHAP waterfall plots for representative individuals across all three risk strata (baseline $E[f(X)] = 0.438$). Bars show signed feature contributions ϕ_j from the baseline to the final predicted susceptibility score.

The low-risk case ($\hat{p} = 0.091$) is the most concentrated: the two crowding index variables alone account for -0.21, and the subsequent protective contributions come exclusively from household-level socioeconomic features (mean DispInk04, income range variables).

4.4.4 XAI Summary

The three-level XAI analysis yields a consistent and epidemiologically grounded account of SARS-CoV-2 household secondary transmission risk across both prediction levels. At the *household level*, the maximum index-case age and mean family income are the two primary global drivers, with household size ranking third. The subgroup analysis reveals that housing type (association apartment) distinguishes the rare high-risk households, while birth-country diversity is a persistent discriminating variable across the moderate-to-low risk strata, pointing to ethnic and cultural heterogeneity as a complex modifier of household transmission dynamics. At the *individual level*, family income and member age dominate globally, but within the high- and low-risk strata the household crowding index emerges as the leading factor, indicating that physical transmission environment is the critical differentiator at the extremes of the susceptibility distribution. Near the classification boundary (moderate stratum), income-based deprivation measures retake the leading role.

Across both levels, the consistent presence of income-related features points to economic vulnerability as a cross-cutting contributor to household transmission risk in this pre-vaccination Swedish cohort, a pattern that aligns with international evidence on social determinants of COVID-19 and that points to socioeconomic targeting as a high-yield strategy for future epidemic preparedness. The crowding index, prominent in the individual-level subgroup and local analyses, adds a spatial dimension to this picture: it is not only the level of household income that matters, but the physical density of the shared living environment. The complementarity of the two prediction levels is precisely the added value of the unified framework: the household model identifies *which* families face elevated transmission risk, while the individual model identifies *who* within those families warrants priority protection.

Chapter 5

LLM Application for Policy-Oriented Explanation

The design of the LLM-based explanation system, including the structured inference report format, the chain-of-thought prompting strategy, and the selection of LLMs evaluated, is presented in Chapter 3, Section 3.4. This chapter introduces the evaluation framework used to assess explanation quality (Section 5.1), presents the empirical results (Section 5.2), and discusses practical deployment considerations (Section 5.3).

5.1 Evaluation Framework

We designed a two-tier evaluation framework combining qualitative expert assessment with objective quantitative metrics computed directly from 10 saperate samples' source `report.txt`. The two tiers capture complementary modes: qualitative metrics detect problems that are imperceptible to automated scoring, such as subtle causal inference or imprecise mechanism descriptions, while quantitative metrics provide reproducible, objective measurements of factual coverage and numerical fidelity. Table 5.1 summarises all metrics before each is described in detail below.

5.1.1 Qualitative Metrics (Part A)

The four qualitative metrics are assessed across each full LLM output and are structured as a *layered dependency chain*: a score of zero on any lower layer caps all subsequent layers at zero. This ordering reflects the principle that logical consistency cannot be meaningfully assessed in incomprehensible text, and that epidemiological plausibility cannot be assessed in logically inconsistent reasoning.

A1 (Comprehensibility) assesses whether the explanation is clearly expressed and understandable to a general reader, independent of whether the content is correct. A score of 2 indicates an unambiguous, well-structured output in which all technical terms are avoided or explained; a score of 1 indicates mostly clear text with some vague or ambiguous statements; and a score of 0 indicates an incomprehensible or self-contradictory output at the level of expression.

Table 5.1: Two-tier LLM evaluation framework. Part A: qualitative metrics scored 0–2 by a human expert and an LLM evaluator. Part B: quantitative metrics computed against the source `report.txt`.

Tier	Metric	Label	Range
Qualitative	Comprehensibility	A1	0–2
Qualitative	Soundness	A2	0–2
Qualitative	Plausibility	A3	0–2
Quantitative	HH Top-2 Recall	B1	0–100%
Quantitative	HH Top-3 Recall	B2	0–100%
Quantitative	IND Top-2 Recall (per member)	B3	0–100%
Quantitative	IND Top-3 Recall (per member)	B4	0–100%
Quantitative	Mean IND Top-2 Recall	B5	0–100%
Quantitative	Mean IND Top-3 Recall	B6	0–100%
Quantitative	Direction Accuracy	B7	0–100%
Quantitative	Faithfulness	B8	0–100%

A2 (Soundness), contingent on Comprehensibility ≥ 1, assesses whether the reasoning is internally consistent and logically valid without requiring specialised domain knowledge to detect errors. It is scored 2 if all reasoning steps are coherent and conclusions follow from stated evidence, 1 if the reasoning contains gaps or unsupported leaps, and 0 if a direct logical contradiction is present. Failure modes specific to this task include claiming that a feature "causes" infection when the prompt specifies it as a model attribution (SHAP $\neq$ causation), inferring prediction correctness from model confidence alone, and applying household-level reasoning to individual-level conclusions or vice versa.

A3 (Plausibility), contingent on Soundness ≥ 1, assesses whether the causal claims and risk-factor interpretations are consistent with established epidemiological knowledge, including the direction, magnitude, and mechanism described. A score of 2 requires that both causal direction and proposed mechanism align with epidemiological consensus; a score of 1 is assigned when the direction is correct but the stated mechanism is imprecise or partially incorrect; and a score of 0 is given when direction or mechanism clearly contradicts domain knowledge. Because this metric requires genuine epidemiological expertise, it is assessed by a qualified domain expert.

The maximum qualitative score per report is 6. Qualitative metrics are assessed independently by a human domain expert.

5.1.2 Quantitative Metrics (Part B)

Quantitative metrics are computed objectively against the SHAP tables in `report.txt`. Feature names in the LLM output may appear as paraphrases (e.g., "homecare alarms" for `TRYGG_1`), so a canonical feature glossary is used to map paraphrases to feature names before scoring; only features explicitly mentioned by name or unambiguous paraphrase are counted. Let HH_top_k denote the set of top-k household SHAP features by |SHAP|, $\text{IND}_{i,\text{top}_k}$ the top-k individual SHAP features for member i, M_{HH} the household features mentioned in the explanation, M_{IND_i} the features mentioned for member i, and N_{mem} the number of non-index members with individual predictions.

B1 and B2 (Household Top-2 and Top-3 Recall) measure the fraction of the most influential household SHAP features explicitly covered anywhere in the explanation:

$$\text{HH\_Top2\_Recall} = \frac{|M_{\text{HH}} \cap \text{HH\_top}_2|}{2}, \quad \text{HH\_Top3\_Recall} = \frac{|M_{\text{HH}} \cap \text{HH\_top}_3|}{3}. \tag{5.1}$$

Pairing Top-2 and Top-3 reveals whether coverage degrades at the third-ranked feature, which would suggest that LLMs tend to focus on only the most dominant signal.

B3 and B4 (Per-Member Individual Top-2 and Top-3 Recall) apply the same principle at the individual level, computed independently for each non-index member i:

$$\text{IND}_{i,\text{Top2}} = \frac{|M_{\text{IND}_i} \cap \text{IND}_{i,\text{top}_2}|}{2}, \qquad \text{IND}_{i,\text{Top3}} = \frac{|M_{\text{IND}_i} \cap \text{IND}_{i,\text{top}_3}|}{3}. \tag{5.2}$$

Reporting these per-member values rather than only an aggregate is essential for detecting uneven coverage across household members.

B5 and B6 (Mean Individual Top-2 and Top-3 Recall) aggregate B3 and B4 across all members:

$$\overline{\text{IND\_Top2}} = \frac{1}{N_{\text{mem}}} \sum_{i=1}^{N_{\text{mem}}} \text{IND}_{i,\text{Top2}}, \quad \overline{\text{IND\_Top3}} = \frac{1}{N_{\text{mem}}} \sum_{i=1}^{N_{\text{mem}}} \text{IND}_{i,\text{Top3}}. \tag{5.3}$$

These summary statistics are reported alongside the per-member values so that uniform coverage can be distinguished from situations where a high mean conceals uneven treatment of individual members.

B7 (Direction Accuracy) measures the proportion of features mentioned in the explanation for which the LLM correctly states the effect direction (risk-increasing vs. risk-decreasing) consistent with the sign of the corresponding SHAP value:

$$\text{Direction\_Accuracy} = \frac{\#\text{ features with correct direction stated}}{\#\text{ features with any explicit direction claim}}. \tag{5.4}$$

Only features for which the explanation makes an explicit directional claim (e.g., "increases risk", "reduces susceptibility", "pushes prediction higher") contribute to

the denominator; qualitative descriptors without directionality (e.g., "important", "notable") are excluded. The SHAP sign convention is that a positive value means the model pushes the score higher towards flagging. The metric is pooled across the household and all individual members.

B8 (Faithfulness) measures the proportion of numerical values explicitly cited in the explanation that exactly match the source `report.txt` within rounding precision:

$$\text{Faithfulness} = \frac{\#\text{ correctly cited values}}{\#\text{ total cited values}}. \tag{5.5}$$

Countable cited values include P(transmission) and P(susceptible) scores, threshold values, raw feature values from the SHAP table (e.g., age, income), and household size or member counts. Qualitative descriptors such as "elevated" or "significantly higher" do not count, nor do paraphrased ranges unless the midpoint matches. A rounding tolerance of ± 1 unit at the reported precision is applied (e.g., "67%" is accepted for a source value of 0.671).

5.2 Results and Analysis

5.2.1 Qualitative Results

All five models achieved the maximum score of 2.00 on metrics A1 (Comprehensibility) and A2 (Soundness) across all ten evaluation runs, yielding perfect scores on these two dimensions without exception. All five models produced outputs that were unambiguously well-structured, free of internal logical contradictions throughout all runs.

The A3 (Plausibility) metric, which assesses whether the causal mechanisms and risk-factor interpretations are epidemiologically coherent, requires domain expertise. For this reason, expert plausibility assessment was conducted on a representative subset of outputs rather than applied exhaustively across all five models and all ten runs. Claude Sonnet 4.5 was selected as the focal model for in-depth qualitative analysis because it demonstrated the most consistently structured and detailed reports across preliminary inspection, making it the most informative candidate for expert review. The outputs of the remaining models (GPT-5.3, Grok 4.2, DeepSeek V3.2, and Llama 4) were verified to achieve perfect scores on A1 and A2, but their A3 plausibility was not subjected to the same depth of independent expert assessment.

Table 5.2 reports the qualitative scores for Claude Sonnet 4.5 across all four metrics. The expert assessor confirmed that the risk-factor interpretations, causal directions, and proposed mechanisms produced by Sonnet were epidemiologically sound and showed no material deficiencies.

These results carry two important implications. First, they confirm the *validity of the pipeline*: the combination of the structured inference report format and the chain-of-thought prompting strategy is sufficient to elicit policy-grade explanations from current frontier LLMs without any model-specific fine-tuning. Second, the perfect scores on A1 and A2 across all five architecturally distinct models indicate that these qualitative properties are robust features of the pipeline rather than ca-

Table 5.2: Mean qualitative scores (A1–A3) for Claude Sonnet 4.5 across 10 evaluation runs (max. 2 per metric, max. 6 total). All other models achieved 2.00/2.00 on A1 and A2; A3 was not assessed by domain expert for those models.

Model	A1 Comp.	A2 Sound.	A3 Plaus.
Claude Sonnet 4.5	2.00	2.00	2.00

pabilities unique to any particular model family, strengthening the generalisability of the approach.

5.2.2 Quantitative Results

Table 5.3 reports the mean quantitative metric scores per LLM. As with the qualitative tier, all five models achieved perfect scores across all quantitative metrics in all ten runs: household and individual feature recall was 100% at both Top-2 and Top-3 thresholds, direction accuracy was 100%, and faithfulness to numerically cited values was 100%.

Table 5.3: Mean quantitative metric scores per LLM across 10 evaluation runs. Recall and accuracy in percentages.

Model	B1 HH-R@2	B2 HH-R@3	B5 IND-R@2	B6 IND-R@3	B7 Dir.	B8 Faith.
Claude Sonnet 4.5	100%	100%	100%	100%	100%	100%
GPT-5.3	100%	100%	100%	100%	100%	100%
Grok 4.2	100%	100%	100%	100%	100%	100%
DeepSeek V3.2	100%	100%	100%	100%	100%	100%
Llama 4	100%	100%	100%	100%	100%	100%

The household and individual feature recall results (B1–B6) confirm that all models accurately identified and discussed all top-ranked SHAP features at both prediction levels, with no tendency to favour household-level content over the more granular per-member individual profiles. Perfect direction accuracy (B7) indicates that the sign conventions embedded in the inference report format, via "HIGHER" and "LOWER" direction labels for each SHAP attribution, were consistently interpreted and correctly reflected in generated text, even for features with indirect epidemiological interpretations such as homecare alarm activations or drug-class indicators. Perfect faithfulness (B8) confirms that no model fabricated or rounded-away numerical values; all cited scores, thresholds, and raw feature values were reproduced with full accuracy within rounding tolerance. The complete set of evaluation reports for all five models across all ten runs, including the full LLM-generated narratives, the structured inference reports, and the per-run metric scores for both qualitative and quantitative tiers, is available in the accompanying thesis repository [44].

5.2.3 Discussion of Results

The uniformly maximal scores under the defined evaluation framework is the central empirical finding of this chapter. It is important to interpret this result correctly.

The evaluation was not designed to rank or differentiate among the selected LLMs; the five models were chosen to represent the breadth of frontier model families, not to identify the best model for this specific task. It constitutes the evidence for the pipeline generalisability of the framework. The ceiling result implies that any capable frontier LLM, given the same structured input and prompting design, is sufficient to produce policy-grade explanations.

Furthermore, these results lay the methodological groundwork for applying the same evaluation framework to more constrained settings. A plausible future direction is to deploy the explanation pipeline using a smaller, domain-specific model, such as a fine-tuned biomedical LLM or a quantised open-source variant suitable for on-premise deployment in a healthcare setting. Such models may not inherit the strong instruction-following capabilities of frontier models, and the evaluation framework developed here can serve as a benchmark to determine whether a candidate smaller model meets the quality thresholds demonstrated in this study, or whether additional prompt engineering or fine-tuning is required to close the gap.

5.3 Deployment Considerations

The perfect evaluation results across all five models have direct implications for practical deployment of the LLM explanation layer.

On **prompt robustness**, qualitative scores observed across all ten runs per model indicate that the current system prompt is highly robust for frontier LLMs. Nevertheless, the prompt was designed with explicit safeguards (negative-example scope constraints, coverage-parity instructions for multi-member households, and a complete feature glossary) that collectively contribute to this stability. These elements should be retained and periodically reviewed if the feature set or model vocabulary evolves.

On **human oversight**, despite the evaluation results, LLM-generated explanations are intended as a communication aid for qualified public health experts rather than a replacement for domain judgment. Households with unusual socioeconomic profiles or atypical combinations of medical history features can still generate suboptimal outputs. Qualified expert review of LLM explanations remains advisable, particularly for high-stakes intervention decisions.

On **version stability**, LLM behaviour is sensitive to model version updates. Production deployments must establish a re-evaluation protocol whenever a model is updated, as prompt compatibility is not guaranteed across versions. Running the same ten-repetition evaluation on any new model or version against the existing evaluation cases is suggested when model version is updated.

Chapter 6

Discussion

6.1 Strengths and Novelty

Several aspects of this work represent contributions relative to the existing literature on machine learning for infectious disease prediction and explainable AI.

Unified multi-level prediction framework. This is a novel ML system applied to population-wide national register data that jointly addresses household-level transmission probability and individual-level susceptibility scoring within a single coherent pipeline. Prior work has addressed one level or the other but not both simultaneously, and the complementary nature of the two outputs, which together answer the "which household?" and "which member?" questions for intervention prioritisation, is a novel contribution to the epidemiological ML literature.

The selection of TabPFN. The selection of TabPFN as the optimal model is motivated not only by empirical performance but by the alignment between its causal generative prior and the structure of epidemiological register data. The model's in-context learning mechanism and its pretraining on SCM-generated synthetic datasets make it a principled choice aligned with real-world data-generating structures. This choice anticipates future work on counterfactual and causal intervention analysis, for which TabPFN's pretraining provides a natural starting point, beyond the associative SHAP explanations reported in this study.

Multi-granularity XAI. The three-level SHAP analysis, covering global population-wide attribution, subgroup stratification by predicted risk stratum, and local instance-level decomposition, provides a more comprehensive picture of model behaviour than single-level explanations. The structured progression from global to local enables epidemiologists to reason about population-level risk factors, subgroup-level policy targets, and individual clinical cases within a unified explanatory framework.

LLM evaluation framework. The two-tier evaluation framework combining layered qualitative metrics (Comprehensibility, Soundness, Plausibility, Scope Adherence) and objective quantitative metrics (feature recall, direction accuracy, faithfulness) is a systematic evaluation protocol designed specifically for LLM-generated narratives of ML-based XAI outputs in a public health context. Its structure is task-agnostic and could be applied in any domain where an LLM is used to translate SHAP attributions into policy-readable language.

6.2 Limitations

Register data constraints. The Swedish national register data, while comprehensive in population coverage, is subject to limitations inherent to routine administrative data. Cases selected in 2020 were imperfect: constrained testing capacity led to under-detection of asymptomatic and mildly symptomatic individuals, biasing household outcome labels toward false negatives. The household definition, specifically co-residence at the same registered address on 1 January 2020, which does not capture temporary co-habitation or address changes during the epidemic. The study is also restricted to the first epidemic wave in 2020, prior to vaccination, variant succession, and widespread rapid testing; the learned feature–outcome relationships may not generalise to later pandemic phases or different pathogens without retraining.

Model scalability and threshold sensitivity. TabPFN's in-context learning imposes a practical upper bound on training context size ($N_{\text{tr}}^{\max} = 45{,}000$), necessitating the subsampling strategy described in Sections 3.2.5. The LLM-generated reports frame risk relative to the operational thresholds τ_{hh}^{*} and τ_{ind}^{*}, presenting them as meaningful boundaries for public health communication. However, these thresholds are decision boundaries derived from F1-score optimisation on a class-imbalanced dataset and are fundamentally engineering constructs: they reflect the point that maximises classification utility under the chosen metric, not any epidemiologically meaningful probability of infection. Future work should address this through post-hoc probability calibration, which would align the model's output scale with empirical infection rates and enable thresholds to be set on clinically or epidemiologically grounded grounds rather than metric-optimisation grounds.

KernelSHAP approximation. KernelSHAP computes Shapley values by marginalising over a background dataset $\mathcal{B}$ sampled from the test set; the resulting attributions are therefore conditional on that specific background choice. In this study, $\mathcal{B}$ was constructed via stratified random sampling from the test set, and this relatively small background size introduces non-negligible approximation error: different background samples can yield materially different SHAP values, particularly for low-contribution features.

LLM version instability and evaluation scale. LLM behaviour is sensitive to model version updates, and perfect evaluation scores observed here cannot be assumed to transfer to future model versions or to smaller, domain-specific models suitable for privacy-constrained on-premise deployment. The Plausibility (A3) metric relied on a single epidemiology expert for human assessment; a larger expert panel would provide more reliable inter-rater reliability estimates.

6.3 Future Work

Integration of causal inference. The most natural extension of the XAI analysis is to move from associative SHAP attributions toward causal effect estimation. TabPFN's causal generative prior provides a principled starting point: future work could apply structural causal models or doubly robust estimation to quantify the causal effect of household income or occupancy density on transmission probabil-

ity, disentangling confounded pathways that SHAP cannot separate. This would strengthen the evidential basis for targeted policy interventions.

Prospective and real-time deployment. The current system is retrospective, trained and evaluated on a fixed historical cohort. Deploying the system in real time would require automating the full pipeline from data to model scoring and LLM narration, and monitoring for performance would degrade as the population and epidemic context shifts over time. Here another important challenge is the legal aspect, particularly the need for privacy-preserving machine learning: any live deployment that processes individual-level health and socioeconomic register data must comply with data protection regulations, which may constrain how model inputs are stored, transmitted, and audited, and motivates the development of techniques such as federated learning or differential privacy to enable inference without centralizing sensitive records.

Evaluation with domain-specific smaller models. The LLM evaluation demonstrated ceiling performance for frontier models. The practically important question for deployment in privacy-constrained healthcare settings is whether smaller, locally deployable models can match this performance. The evaluation framework developed here provides a ready-made benchmark, and the inference report format and prompting strategy could be adapted for fine-tuning on domain-specific data.

Extension to other pathogens and settings. The pipeline architecture, covering register-based feature extraction, multi-level ML prediction, KernelSHAP attribution, LLM narration, is disease-agnostic in design. Applying it to other respiratory pathogens (influenza) or to pathogens with strong household clustering (norovirus, tuberculosis) would test the framework's generalisability and generate comparative epidemiological evidence on how household risk factor profiles differ across diseases.

Chapter 7
Conclusion

This thesis set out to address a structural gap at the intersection of machine learning, public health, and human-AI communication: the absence of an integrated, fully explainable system that predicts SARS-CoV-2 household secondary transmission risk at multiple levels of granularity and translates its outputs into plain language accessible to non-technical decision-makers. The system developed here aims to close this gap by combining a unified multi-level ML prediction framework, a structured three-level XAI analysis, and an LLM-based explanation layer, applied to linked Swedish COVID-19 register data from the 2020 pre-vaccination epidemic wave.

Four classifiers were benchmarked at both prediction levels, i.e., individual and household-level. TabPFN achieves the best AUC-ROC and Macro F1 at the household level (AUC-ROC: 0.645; Macro F1: 0.568), and the highest Macro F1 at the individual level (Macro F1: 0.590; AUC-ROC: 0.717). At the individual level, XGBoost achieves a slightly higher AUC-ROC (0.726), but TabPFN's advantage in Macro F1 makes it the preferred model for balanced sensitivity-specificity performance across both classes. In the context of infectious disease surveillance, where the balance between correctly flagging high-risk household members and avoiding unnecessary resource expenditure on low-risk ones is critical, Macro F1 is a more directly actionable summary criterion, and TabPFN leads on this measure at both levels. Beyond its empirical performance, TabPFN's causal generative prior positions it as a natural foundation for future causal inference work that goes beyond the associative explanations currently provided by SHAP.

The three-level SHAP analysis identifies socioeconomic deprivation and age structure as key risk factors at both prediction levels, with distinct leading features at each. At the household level, the age of the index case is the dominant predictor, with family income as the second-ranked driver and household size ranking third; at the individual level, income and age are the leading predictors. The consistent presence of income-related features across both levels suggests an association between economic vulnerability and SARS-CoV-2 within-household spread in the Swedish pre-vaccination population, though it is important to note that SHAP values reflect the model's learned predictive associations rather than causal relationships; whether income acts as a direct determinant or as a proxy for unmeasured confounders such as housing density, occupational exposure, or health literacy cannot be established from these analyses alone. With this caveat in mind, the findings are epidemiologically coherent and internally consistent across global

and subgroup analyses, and they are suggestive for policy prioritisation: they point toward socioeconomically deprived households with older index cases as a plausible target for preventive intervention, pending further causal investigation.

The LLM evaluation demonstrates that the complete pipeline from raw data to policymaker-readable narrative is both functional and robust. All five frontier models evaluated, spanning proprietary and open-source architectural families, achieved perfect scores on the novel two-tier evaluation framework across ten independent runs, with the supervising epidemiology expert confirming the epidemiological soundness of the evaluated representative outputs. The uniformity of this result across model families is a key practical finding: this consistency across model families suggests that the pipeline design, rather than the choice of any specific LLM, is the primary driver of explanation quality, and that practitioners can therefore select a model based on practical considerations such as cost, latency, or data privacy requirements.

The broader significance of this work is methodological as much as epidemiological. The combination of a causally motivated tabular ML model, a multi-granularity post-hoc explanation framework, and an LLM-based narration layer constitutes a template for how complex predictive systems can be made simultaneously transparent to engineers and interpretable to domain practitioners. Applied to infectious disease surveillance, this pipeline template enables the kind of evidence-based, rapidly deployable, human-readable risk communication that epidemic preparedness demands. Applied more broadly, it illustrates a path toward AI systems that are not only more accurate than conventional models and interpretable, but genuinely communicable, one of the prerequisites for their responsible deployment in high-stakes public health settings.

References

[1] Qi-Lin Jing, Ming-Jin Liu, Zhi-Biao Zhang, et al. Household secondary attack rate of COVID-19 and associated determinants in Guangzhou, China: a retrospective cohort study. *The Lancet Infectious Diseases*, 20(10):1141–1150, 2020. doi: 10.1016/S1473-3099(20)30471-0.

[2] Zachary J. Madewell, Wan Wee Yang, Ira M. Longini, et al. Factors associated with household transmission of SARS-CoV-2: An updated systematic review and meta-analysis. *JAMA Network Open*, 4(8):e2122240, 2021. doi: 10.1001/jamanetworkopen.2021.22240.

[3] Hannah F. Fung, Leonardo Martinez, Fernando Alarid-Escudero, Joshua A. Salomon, David M. Studdert, Jason R. Andrews, and Jeremy D. Goldhaber-Fiebert. The household secondary attack rate of SARS-CoV-2: a rapid review. *Clinical Infectious Diseases*, 73(S2):S138–S145, 2021. doi: 10.1093/cid/ciaa1558.

[4] World Health Organization. WHO coronavirus (COVID-19) dashboard. *World Health Organization*, 2020. URL `https://covid19.who.int/`. Accessed 2024.

[5] Wei Li, Bo Zhang, Jie Lu, Shiqi Liu, Zhenghai Chang, Cao Peng, Xiaohan Liu, Pinyuan Zhang, Ying Liu, Yanping Hu, and Jun Wang. Characteristics of household transmission of COVID-19. *Clinical Infectious Diseases*, 71(8):1943–1946, 2020. doi: 10.1093/cid/ciaa450.

[6] William Ogilvy Kermack and Anderson Gray McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927. doi: 10.1098/rspa.1927.0118.

[7] Zachary J. Madewell, Yang Yang, Ira M. Longini, M. Elizabeth Halloran, and Natalie E. Dean. Household transmission of SARS-CoV-2: a systematic review and meta-analysis. *JAMA Network Open*, 3(12):e2031756, 2020. doi: 10.1001/jamanetworkopen.2020.31756.

[8] Fang Li, Yan-Yan Li, Ming-Jin Liu, et al. Household transmission of SARS-CoV-2 and risk factors for susceptibility and infectivity in Wuhan: a retro-

spective observational study. *The Lancet Infectious Diseases*, 21(5):617–628, 2021. doi: 10.1016/S1473-3099(20)30981-6.

[9] Naja Hulvej Rod, Alex Broadbent, Morten Hulvej Rod, et al. Complexity in epidemiology and public health. Addressing complex health problems through a mix of epidemiologic methods and data. *Epidemiology (Cambridge, Mass.)*, 34(4):505–514, 2023. doi: 10.1097/EDE.0000000000001612.

[10] Martyn Fyles, Elizabeth Fearon, Christopher Overton, Tom Wingfield, Graham F. Medley, Ian Hall, Lorenzo Pellis, and Thomas House. Using a household-structured branching process to analyse contact tracing in the SARS-CoV-2 pandemic. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1829):20200267, 2021. doi: 10.1098/rstb.2020.0267.

[11] Alberto Aleta, David Martín-Corral, Ana Pastore y Piontti, Marco Ajelli, Maria Litvinova, Matteo Chinazzi, Natalie E. Dean, M. Elizabeth Halloran, Ira M. Longini, Stefano Merler, Alex Pentland, Alessandro Vespignani, Esteban Moro, and Yamir Moreno. Modelling the impact of testing, contact tracing and household quarantine on second waves of COVID-19. *Nature Human Behaviour*, 4(9):964–971, 2020. doi: 10.1038/s41562-020-0931-9.

[12] Mingxuan Liu, Yifan Liu, and Jing Liu. Machine learning for infectious disease risk prediction: a survey. *ACM Computing Surveys*, 57(7):1–37, 2025. doi: 10.1145/3719663.

[13] Kevin W. K. Yang, Catriona F. Paris, Kathryn T. Gorman, et al. Factors associated with resistance to SARS-CoV-2 infection discovered using large-scale medical record data and machine learning. *PLOS ONE*, 18(2):e0278466, 2023. doi: 10.1371/journal.pone.0278466.

[14] Auriel A. Willette, Stephanie A. Willette, Qian Wang, et al. Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study. *Scientific Reports*, 12:7736, 2022. doi: 10.1038/s41598-022-07307-z.

[15] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations 2023*, 2023.

[16] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 01 2025. doi: 10.1038/s41586-024-08328-6. URL `https://www.nature.com/articles/s41586-024-08328-6`.

[17] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[18] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.

[19] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017.

[20] Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games*, volume II, pages 307–317. Princeton University Press, Princeton, NJ, 1953. doi: 10.1515/9781400881970-018.

[21] Johannes Allgaier, Laura Mulansky, Rachel L. Draelos, et al. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artificial Intelligence in Medicine*, 143:102616, 2023. doi: 10.1016/j.artmed.2023.102616.

[22] Krishnaraj Chadaga, Srikanth Prabhu, Niranjana Sampathila, et al. Explainable artificial intelligence approaches for COVID-19 prognosis prediction using clinical markers. *Scientific Reports*, 14:1725, 2024. doi: 10.1038/s41598-024-52428-2.

[23] Moosa Tatar, Mohammad Reza Faraji, and Fernando A. Wilson. Social vulnerability and initial COVID-19 community spread in the US South: a machine learning approach. *BMJ Health & Care Informatics*, 30(1):e100703, 2023. doi: 10.1136/bmjhci-2022-100703.

[24] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[25] Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. LLMs for XAI: Future directions for explaining explanations. In *Proceedings of the ACM CHI Workshop on Human-Centered Explainable AI (HCXAI)*, 2024. doi: 10.48550/arXiv.2405.06064.

[26] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, 5(8):873–883, 2023. doi: 10.1038/s42256-023-00692-8.

[27] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[28] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, New York, 2016. ACM. doi: 10.1145/2939672.2939785.

[29] Marwa El Baldi, Zakia Marsou, Fatima Zahra Chellat, Mohamed Chakib Benjelloun, Mohammed Faouzi Belahsen, Tarik Sqalli Houssaini, Khalid Lahmadi, and Karima El Rhazi. Household transmission of SARS-CoV-2 during the first wave of the COVID-19 pandemic: an analysis of secondary attack rates in Moroccan households. *BMC Infectious Diseases*, 25:1022, 2025. doi: 10.1186/s12879-025-11464-7.

[30] Georgia Sovatzidi, Georgia Triantafyllou, George Dimas, et al. Risk assessment of COVID-19 transmission on cruise ships using fuzzy rules. In *Artificial Intelligence Applications and Innovations (AIAI 2024)*, volume 713 of *IFIP Advances in Information and Communication Technology*, pages 336–348, Cham, 2024. Springer. doi: 10.1007/978-3-031-63219-8_25.

[31] Sercan Ö. Arik and Tomas Pfister. TabNet: Attentive interpretable tabular learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021. doi: 10.1609/aaai.v35i8.16826.

[32] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18761–18771, 2021.

[33] Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, et al. Tabpfn-2.5: Advancing the state of the art in tabular foundation models, 2025. URL `https://arxiv.org/abs/2511.08667`.

[34] David Rundel, Lukas Schmid, Bernd Bischl, et al. Interpretability of in-context learning for tabular data: Shapley values and leave-one-covariate-out for TabPFN. *Transactions on Machine Learning Research*, 2024.

[35] Lawrence A. Shaktah, Marco Gustav, Tim Lenz, Junhao Liang, Lars Hilgers, Zunamys I. Carrero, and Jakob Nikolas Kather. Established machine learning matches tabular foundation models in clinical predictions. *medRxiv*, 2026. doi: 10.64898/2026.02.02.26345274. Preprint, posted 4 February 2026.

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, New York, 2016. ACM. doi: 10.1145/2939672.2939778.

[37] Shivani Dhiman, Anjali Thukral, and Punam Bedi. An agentic ai system for disease diagnosis with explanations. *Informatics and Health*, 3(1):32–40, 2026. ISSN 2949-9534. doi: https://doi.org/10.1016/j.infoh.2026.01.001.

[38] Ayoub Belouadah, Marcelo Luis Ruiz-Rodríguez, Sylvain Kubler, and Yves Le Traon. Evaluating the effectiveness of llms for explainable deep reinforcement learning. *Machine Learning with Applications*, 22:100795, 2025. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2025.100795.

[39] World Health Organization. Household transmission investigation protocol for 2019-novel coronavirus (COVID-19) infection. World Health Organization, 2020. URL https://www.who.int/publications/i/item/household-transmission-investigation-protocol-for-2019-novel-coronavirus-(2019-ncov)-infection.

[40] UK Health Security Agency. COVID-19 infectious period: a rapid evidence review (update 1). Technical report, UK Health Security Agency, 2023. URL https://assets.publishing.service.gov.uk/media/658580c723b70a0013234e62/COVID-19-infectious-period-a-rapid-evidence-review-_update-1_.pdf. Publishing reference: GOV-15887.

[41] SWECOV. SWECOV: Swedish register-based research project on COVID-19. multidisciplinary research collaboration using quantitative methods and comprehensive register data about the whole Swedish population. https://swecov.se. Accessed December 31, 2025.

[42] A. S. Hashemi, Y. Litins'ka, J. Björk, et al. Explainable and just AI in data-driven disease surveillance. Active project from 2025-06-01 to 2027-05-31 at Lund University. https://portal.research.lu.se/en/projects/explainable-and-just-ai-in-data-driven-disease-surveillance/, 2025.

[43] UPPMAX, Uppsala University. Bianca — sensitive data hpc cluster. https://www.uu.se/en/centre/uppmax, 2025. NAISS, Uppsala Multidisciplinary Center for Advanced Computational Science.

[44] Mingtong Du. An explainable ai and llm-assisted framework for within-household sars-cov-2 secondary transmission risk prediction in sweden. Master's thesis, Lund University, 2025.

Printed by Tryckeriet i E-huset, Lund 2026

LUND
UNIVERSITY

Series of Master's theses
Department of Electrical and Information Technology
LU/LTH-EIT 2026-1136
http://www.eit.lth.se