CHAPTER V

DISCUSSION AND CONCLUSION

5.1 Discussion

This research was undertaken with the primary objective *to develop techniques for determining model parameters in Double Generalized Additive Models (DGAMs)* using Smoothing Splines, Explainable Boosting Machine (EBM), Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Regression (SVR). In DGAMs, model development must address both the conditional mean and the conditional dispersion. Therefore, the ability to learn nonlinear effects while accommodating heteroscedasticity is central to method selection and to the reliability of uncertainty quantification.

Overall, the empirical findings in Phase II support the motivation of DGAM for the present problem setting. In many experimental conditions, DGAM improved distribution-focused criteria such as negative log-likelihood (NLL), continuous ranked probability score (CRPS), and Akaike information criterion (AIC). At the same time, point-prediction metrics such as MAE and RMSE were often comparable to GAM. This pattern is consistent with the DGAM design: the primary benefit is improved modeling of the predictive distribution via a dispersion submodel, rather than guaranteed improvement in point estimates.

5.1.1 Smoothing Splines for DGAM

Smoothing splines provide a principled baseline for DGAMs because they estimate smooth nonlinear functions without requiring a prespecified parametric form. Within the DGAM framework, splines naturally support separate smooth components for the mean and dispersion submodels, allowing the analyst to directly interpret how predictors influence both central tendency and variability. A key advantage is the explicit control of model complexity through smoothing parameters, which makes the bias–variance trade-off transparent and supports statistically grounded selection via cross-validation or infor-

mation criteria.

In addition, spline-based DGAMs retain strong inferential tooling, including approximate standard errors, confidence bands for smooth effects, and hypothesis tests for functional terms. This remains particularly valuable in scientific and health analytics applications where interpretability is as important as predictive accuracy and where communicating uncertainty requires clear model explanations.

5.1.2 Explainable Boosting Machine (EBM) for DGAM

EBM offers an attractive compromise between interpretability and predictive strength by learning an additive structure composed of data-driven shape functions (and, when needed, limited interactions). In a DGAM setting, this additive decomposition maps well to distributional regression because both the mean and dispersion components can be modeled as sums of transparent effects. This supports a practical workflow in which the analyst can examine which predictors drive the mean response versus which predictors primarily drive predictive uncertainty.

A further advantage is that EBM can capture nonlinear patterns without requiring extensive manual feature engineering. However, as with any flexible learner, it is important to validate performance using both probabilistic metrics (e.g., NLL and CRPS) and calibration diagnostics, particularly when EBM is used for the dispersion component.

5.1.3 XGBoost for DGAM

XGBoost is a powerful learner for capturing complex nonlinearities and feature interactions. When integrated into DGAM, XGBoost can be especially effective for the dispersion submodel in settings where the variance structure changes in a non-smooth or interaction-driven manner. This flexibility can translate into strong likelihood-based improvements when heteroscedasticity is pronounced and cannot be adequately captured by simpler smoothers.

Nevertheless, the main practical challenge with XGBoost lies in hyperparameter tuning and the risk of overfitting, particularly in the dispersion model. Overfitting the dispersion component can degrade out-of-sample calibration even if in-sample fit appears

improved. For this reason, XGBoost-based DGAM models should be accompanied by careful validation across splits and, ideally, by regularization strategies that prioritize stable uncertainty estimates.

5.1.4 Random Forest (RF) for DGAM

Random Forest provides a robust nonparametric approach that can learn nonlinear relationships and interactions with relatively limited tuning. In DGAM, RF can serve as an effective dispersion learner because it can adapt to complex variance patterns that arise in real-world healthcare and operational data. A practical strength of RF is its stability: compared with boosting approaches, RF tends to be less sensitive to hyperparameter choices and less prone to overly sharp variance estimates.

However, RF interpretability is typically lower than spline-based or EBM-based approaches. Therefore, RF is often most suitable when the primary objective is improving distributional fit and calibration rather than maximizing interpretability of individual functional effects.

5.1.5 Support Vector Regression (SVR) for DGAM

SVR is a competitive technique when the underlying signal is nonlinear but relatively smooth, and when regularization is important for generalization. In the DGAM framework, SVR can be used as a dispersion learner to capture systematic changes in residual variability driven by covariates. Its kernel-based formulation provides flexibility while retaining explicit control over model complexity through regularization parameters.

The main limitation of SVR is computational scalability and sensitivity to kernel and parameter settings, particularly for larger datasets. Despite this limitation, SVR remains a useful option in DGAM, especially when the dataset size is moderate and the dispersion structure does not require highly complex interaction modeling.

5.1.6 Recommendations

Based on the empirical findings, DGAM is recommended over a mean-only GAM when heteroscedasticity is present and when reliable uncertainty quantification is required. In such cases, performance should be assessed using probabilistic fit metrics such as NLL and CRPS, as well as calibration diagnostics, rather than relying solely on point-prediction errors. When interpretability is a priority, smoothing splines and EBM provide transparent and communicable models. When the primary goal is flexible capture of variance patterns, RF and XGBoost are strong candidates, with SVR serving as an effective alternative in moderate-sized settings.

Implementation Considerations

- **Model selection should be distribution-aware.** When DGAM is used, selecting models solely by MAE/RMSE can miss the core advantage of dispersion learning. Prefer NLL/CRPS as primary selection criteria and always report calibration (e.g., PICP95).

- **Regularize the dispersion learner explicitly.** Dispersion models can overfit even when the mean model is well-behaved. Use early stopping and/or conservative tree depth for boosting methods, and prefer stable defaults (e.g., RF) when data are limited.

- **Check calibration stability across splits.** Validate that gains in NLL/CRPS persist across train/validation/test ratios and are not driven by a single split configuration.

- **Monitor numerical stability and constraints.** Ensure the dispersion parameterization enforces positivity (e.g., log-link for variance/scale) and verify that predicted scales are within plausible ranges to avoid unstable likelihood values.

- **Interpretability depends on the learner choice.** If interpretability is required, prefer spline- or EBM-based components and use their effect plots for communicating how predictors influence both mean and uncertainty.

Future Research Directions

- Develop automated, principled hyperparameter selection to make DGAM training consistent across learners.
- Improve uncertainty quantification for ML-based DGAMs, including calibrated predictive intervals and reliability diagnostics.
- Extend the DGAM framework to additional exponential-family responses and link functions beyond the Gaussian case.
- Compare DGAM with GAMLSS (location–scale–shape modeling) to capture skewness and heavy tails when dispersion alone is insufficient.
- Integrate deep learning components for mean and dispersion to model high-dimensional, nonlinear effects more flexibly.
- Provide robust, well-documented software packages implementing these methods in widely used statistical environments.
- Conduct large-scale simulation studies to characterize finite-sample behavior, calibration, and failure modes under controlled settings.

5.2 Conclusion

This research has successfully accomplished its primary objective of developing comprehensive techniques for parameter determination in Double Generalized Additive Models using five distinct machine learning approaches. The theoretical foundations, algorithmic specifications, and evaluation across multiple datasets provide a solid basis for both immediate practical application and future methodological development. The integration of Smoothing Splines, Support Vector Regression, Explainable Boosting Machine, Extreme Gradient Boosting, and Random Forest within the DGAM framework demonstrates that diverse machine learning techniques can be systematically adapted to distributional

regression while preserving their individual strengths. This unification of classical statistical modeling with modern machine learning represents a promising direction for the continued evolution of statistical methodology.

As data complexity continues to increase across scientific disciplines, the need for flexible methods that can model both central tendency and variability will only grow. The DGAM framework and associated techniques developed here provide researchers and practitioners with powerful tools to meet this challenge, enabling more accurate predictions, deeper insights, and better-informed decisions in applications ranging from health sciences to environmental modeling and beyond. The journey from theoretical concept through algorithmic development to practical implementation demonstrates the value of rigorous mathematical foundations combined with computational innovation. This research establishes a comprehensive foundation for DGAM methodology that we hope will stimulate further developments and widespread adoption in statistical practice.