Les points de vue identifiés :

**Limitations**

**Portée des Dimensions de Qualité :**

L'étude se concentre sur six dimensions spécifiques de la qualité des données (représentation cohérente, complétude, exactitude des caractéristiques, exactitude des cibles, unicité et équilibre des classes cibles). Il peut y avoir d'autres dimensions de la qualité des données non considérées dans cette étude qui pourraient avoir un impact sur la performance des modèles.

Extension des Dimensions de Qualité : Les recherches futures pourraient explorer des dimensions supplémentaires de la qualité des données, telles que la rapidité, la pertinence et l'interprétabilité, et leur impact sur les modèles de machine learning.  
  
**Pollution Synthétique des Données :**

Les expériences utilisent des méthodes synthétiques pour introduire des erreurs dans les données, ce qui pourrait ne pas capturer toutes les complexités et nuances des problèmes de qualité des données dans le monde réel. Les données réelles peuvent avoir des erreurs plus complexes et interconnectées qui sont difficiles à simuler avec précision.

Application aux Données Réelles : Mener des études similaires avec des ensembles de données réels provenant de différents domaines pourrait fournir plus d'informations sur les implications pratiques des problèmes de qualité des données. Cela pourrait inclure des collaborations avec des partenaires industriels pour accéder à diverses sources de données.  
  
**Sélection des Algorithmes :**

L'étude examine un ensemble spécifique de quinze algorithmes de machine learning. Bien que ces algorithmes couvrent une gamme de méthodes (classification, régression, clustering), il existe de nombreux autres algorithmes qui n'ont pas été inclus. Les résultats peuvent ne pas être généralisables à toutes les méthodes de machine learning possibles.

Élargissement de la Gamme d'Algorithmes : Inclure une variété plus large d'algorithmes de machine learning, y compris des méthodes plus récentes et avancées, pourrait aider à généraliser les résultats. De plus, explorer l'impact sur les méthodes d'ensemble et les modèles hybrides pourrait être précieux.

**Analyse Statique :**

L'étude fournit une analyse statique de l'impact de la qualité des données sur la performance des modèles. Elle ne prend pas en compte les aspects dynamiques de la qualité des données, tels que les changements au fil du temps et leurs effets sur les processus de réentraînement et de mise à jour des modèles.

Études Longitudinales : Des études longitudinales qui suivent l'impact de la qualité des données au fil du temps pourraient fournir une compréhension plus approfondie de la façon dont la dégradation des données affecte la performance des modèles et comment atténuer ces effets par une surveillance et une maintenance continues.  
  
**Outils Automatisés de Qualité des Données :**

Développer et évaluer des outils automatisés pour l'évaluation et l'amélioration de la qualité des données dans les pipelines de machine learning pourrait améliorer l'application pratique des résultats de recherche. Cela inclut l'intégration de contrôles de qualité des données dans les frameworks MLops.

Lien :